Translating AI to Impact: Uncertainty and Human-Agent Interactions in Multi-Agent Systems for Public Health and Conservation

a dissertation presented
by
Elizabeth Carolyn Bondi-Kelly
to
The Department of Computer Science

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Computer Science

Harvard University
Cambridge, Massachusetts
May 2022

Thesis advisor: Professor Milind Tambe Elizabeth Carolyn Bondi-Kelly

Translating AI to Impact: Uncertainty and Human-Agent Interactions in Multi-Agent Systems for Public Health and Conservation

Abstract

Artificial intelligence (AI) is now being applied widely in society, including to support decision-making in important, resource-constrained efforts in conservation and public health. Such real-world use cases introduce new challenges, like noisy, limited data and human-in-the-loop decision-making. I show that ignoring these challenges can lead to suboptimal results in AI for social impact systems. For example, previous research has modeled illegal wildlife poaching using a defender-adversary security game with signaling to better allocate scarce conservation resources. However, this work has not considered detection uncertainty arising from noisy, limited data. In contrast, my work addresses uncertainty beginning in the data analysis stage, through to the higher-level reasoning stage of defender-adversary security games with signaling. I introduce novel techniques, such as additional randomized signaling in the security game, to handle uncertainty appropriately, thereby reducing losses to the defender. I show similar reasoning is important in public health, where we would like to predict disease prevalence with few ground truth samples in order to better inform policy, such as optimizing resource allocation. In addition to modeling such real-world efforts holistically, we must also work with all stakeholders in this research, including by making our field more inclusive through efforts like my nonprofit, Try AI.

Contents

Author List

The following authors contributed to

- Chapter 1: Hoon Oh, Haifeng Xu, Fei Fang, Bistra Dilkina, Milind Tambe.
- Chapter 2: Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, Krishnamurthy (Dj) Dvijotham. This work was conducted while I was an intern at DeepMind. This work was not conducted as a Harvard affiliate.
- Chapter 3: Fei Fang, Debarun Kar, Venil Noronha, Donnabell Dmello, Milind Tambe, Arvind Iyer, Robert Hannaford.
- Chapter 4: Fei Fang, Mark Hamilton, Debarun Kar, Donnabell Dmello, Jongmoo Choi, Robert Hannaford, Arvind Iyer, Lucas Joppa, Milind Tambe, Ram Nevatia.
- Chapter 5: Debadeepta Dey, Ashish Kapoor, Jim Piavis, Shital Shah, Fei Fang, Bistra Dilkina, Robert Hannaford, Arvind Iyer, Lucas Joppa, Milind Tambe.
- Chapter 6: Raghav Jain, Palash Aggrawal, Saket Anand, Robert Hannaford, Ashish Kapoor, Jim Piavis, Shital Shah, Lucas Joppa, Bistra Dilkina, Milind Tambe.
- Chapter 7: Haipeng Chen, Christopher D. Golden, Nikhil Behari, Milind Tambe.
- Chapter 8: Lily Xu, Diana Acosta-Navas, Jackson A. Killian. Equal contribution by Lily Xu. Order determined by coin flip.

Listing of figures

"Together, we must find the best solutions to ensure that the development of AI is an opportunity for humanity, as it is our generation's responsibility to pass down to the next a society that is more just, more peaceful and more prosperous."

~ Audrey Azoulay, Director-General of UNESCO

Acknowledgments

Thank you to everyone who has helped me along my journey so far. First, thank you to my adviser, Milind Tambe, for the many intriguing, inspirational discussions over the past six years. I will never forget our first meeting during PhD visit days in March 2016. I had a wonderful time discussing research ideas and learning about the lab, but I was truly impressed by your ability to make me feel welcome and supported. I also likely would not have started Try AI without our discussions since then. All of your support, advice, and encouragement have played an enormous role in my growth as a scientist and as a person, and I cannot wait to stay in touch and to pass on your wisdom to my own mentees.

Thank you to my dissertation committee members, Yiling Chen, Christopher Golden, and Fernanda Viegas, for your support and guidance with both this dissertation and with career advice in general. Thank you to my qualifying exam committee members at the University of Southern California, Gaurav Sukhatme, Bistra Dilkina, Phebe Vayanos, and Ning Wang, for your support and advice when I was starting to establish my research directions. Thank you to the many mentors and teachers I've had the pleasure of working with throughout my career, for being generous with your wisdom and helping me reach this point.

Thank you to my amazing collaborators, many of whom are included in the Author List. Without you, none of this work would have moved towards impact. Thank you especially to Alexis Stokes for your support for Try AI from the beginning, amongst all of the other vital diversity, equity, inclusion, and belonging initiatives you have led. I cannot express my appreciation enough for all of the emails and discussions on what Try AI should say or become, and hope to work with you on this for many years to come.

Thank you to my labmates and peers for being wonderful friends, asking and answering tough questions, helping me print my dissertation at the last minute, commiserating when things don't work out, and inspiring me to keep innovating.

Thank you to my amazing mentees for your brilliance and enthusiasm. It has been a joy to pass on tips and ideas, but mostly to learn from and collaborate with you.

Last but far from least, thank you to my family. My husband, Jonathan Kelly, has been a constant source of love and encouragement over the past 12 years. He has helped me tackle challenges from wording an email, to making major career choices, to preparing meals so I don't forget to eat during the busy times, to designing the logo for Try AI and many of my presentation templates over the years. He has inspired me with his drive and dedication to our goals. Most of all, he has been there throughout both my successes and disappointments, through happiness and sadness. I'm sure I would not be where I am today without his support. My mother has been a source of inspiration to me my entire life, from her dedication to and joyful interactions with her students, to her grace and intelligence as a professional and as a parent. Growing up watching her example was a huge influence to my career choices, and her wonderful advice didn't hurt either. My father built me a small computer game when I was little, and encouraged my interest in computers from the beginning, including by bringing me a flier for the internship that started me on my career path. Now, I enjoy discussing the latest GPU or new technology with him. Finally, my sister, for her steadfast encouragement in all aspects of my life, and constant friendship even when I fall off of texting for a month or two because of another deadline. I will never have enough words or room to express my gratitude to you all.

0
Introduction

I believe that artificial intelligence (AI) has enormous potential to positively impact the world. In my work, I have endeavored to actualize this potential through use-inspired research in the area of AI for social impact, particularly in the domains of conservation and public health.

As a vital part of these endeavors, I have sought partnerships with stakeholders in both domains. First, in public health, I have studied micronutrient deficiency in Madagascar in collaboration with public health researchers and a public health and environmental research organization in Madagascar called MAHERY[194]. Second, in conservation, I have worked towards preventing illegal wildlife poaching using conservation drones in collaboration with a non-governmental organization called Air Shepherd[6]. I have also created a nonprofit called Try AI[296], devoted to increasing diversity and inclusion in AI through educational opportunities for students, especially those from historically underrepresented groups, to explore AI and its role in society.

Based on these experiences, I have found that AI for social impact often consists of a cycle with three components, as illustrated in Fig. 1:

- Gathering and Analyzing Real-World Data: When collecting and analyzing *real-world* data to better understand a given situation, challenges frequently arise such as noise and limited data. I have developed methods to collect and analyze real-world data, particularly remotely-sensed imagery, in the presence of these challenges.
- Multi-agent Reasoning: Multi-agent interventions often occur after data are gathered and analyzed, for example to deploy limited resources and/or interventions. I have built on the above data and techniques to better inform such multi-agent interventions.
- Deployment: Finally, I have built systems to support deployment in collaboration with stakeholders and domain experts, and have conducted research on the ethics of developing and deploying AI systems.

Through my work in each of these components, I have observed at least two key challenges. First, for an AI system to perform as expected, modeling uncertainty in all components of these systems is needed, as opposed to prior work which may ignore it by focusing on only one component, e.g., multi-agent reasoning. Second, human-agent interaction is likely to occur in each component, particularly because developing and deploying AI systems requires interdisciplinary, diverse, and inclusive teams to truly achieve AI for *positive* social impact.

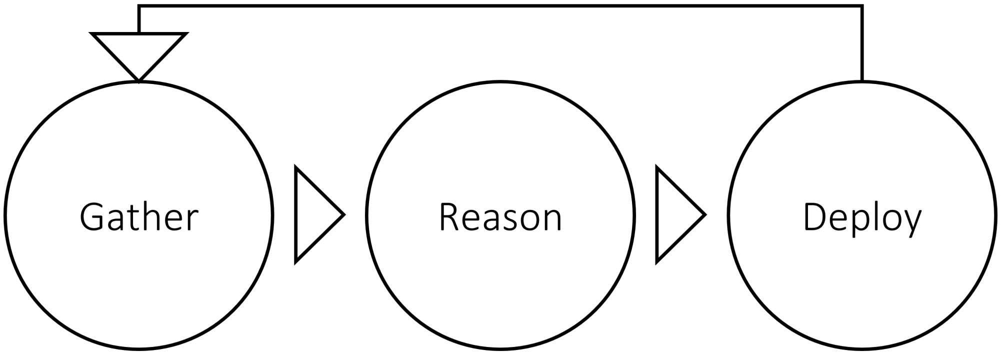


Figure 1: AI for social impact cycle. It involves gathering and analyzing real-world data, then using multi-agent reasoning for recommendations, such as resource allocation, and finally deploying, which may lead to more data collection and analysis.

0.1 Problem Statement and Contributions

In this thesis, I therefore seek to address the question:

How can we account for the challenges of uncertainty and human-agent interactions throughout the AI for social impact pipeline?

My key contributions towards addressing these challenges are my (i) algorithms to account for uncertainty in security games with signaling, (ii) demonstration of the impact of model predictions and uncertainty on human decision-making, (iii) multiple data augmentation strategies to handle limited data, and (iv) field deployment of an algorithm for conservation drones.

0.2 Summary

0.2.1 AI systems for uncertainty throughout pipeline

I will begin a more detailed summary of these contributions by focusing on the conservation domain, and specifically, when we are trying to protect animals from illegal wildlife poaching. Thermal infrared drone imagery can provide real-time updates to park rangers at night to prevent illegal wildlife poaching when it typically occurs. However, such a program is difficult to implement due to the limited resources available to protect enormous parks.

In prior work, this challenge of protecting a national park from illegal poaching with limited resources has been modeled using *game theory*, specifically *security games*, to reason about the interactions between park rangers and poachers. Sensors and drones have also been considered, including with a concept called strategic signaling. During a signal, onboard drone lights can be turned on in hopes of deterring poaching. However, given uncertainty in the domain, whether from the difficulty in detecting humans and animals in thermal drone data or in poachers seeing signals, I showed that ignoring uncertainty in strategic signaling, as in prior frameworks, led to large losses for the park rangers, and therefore (i) introduced a novel reaction stage to the game model, in which a park ranger can visit another target if nothing is observed at first, and (ii) constructed a new signaling scheme which includes signaling when nothing is observed automatically[42]. In fact, this signaling scheme *exploits an informational advantage*, in which the rangers know if there is truly a detection, while the poacher is not sure whether there is truly a ranger responding. I accelerated solving the resulting linear programs by designing a branch and price algorithm, with key novelties in the design of our bound relaxation and MILP secondary problem for column generation. This highlights the importance of accounting for uncertainty from the first stage of the pipeline while doing multi-agent reasoning.

0.2.2 Human-agent interaction to address uncertainty

Furthermore, in order to initiate responses like signaling or rangers checking on targets, this needs to be communicated with park rangers once we've detected a human or animal. In prior work on selective prediction, a method which makes predictions but defers to humans on uncertain images, evaluation has typically been done by simulating human behavior on deferred images, e.g., using historical labels. However, there is growing evidence that humans may be biased in their interactions with AI. My collaborators and I consequently evaluated selective prediction for real-world camera trap images with 200 participants on Prolific[41], and found that varying messages led to a statistically significant impact on human accuracy, supporting the idea that these interactions must be studied carefully. Here, for example, providing a deferral message led to improved human accuracy, e.g., by leading the human decision-maker to concentrate and find a hidden animal, while providing a prediction message proved misleading on occasion and harmed performance. Therefore, if we carefully present AI model information to humans, they may be able to help us overcome AI model uncertainty.

0.2.3 Data augmentation to address uncertainty

Uncertainty may also be addressed via data augmentation. As an example, thermal, aerial imagery which may be used to prevent illegal wildlife poaching is difficult to review in real time, especially all night. Unfortunately, it is difficult to automatically detect humans and animals as well, largely due to noise and object size. Prior work in automatic detection uses traditional computer vision, or other data modalities altogether, which are not available in this case. After testing prior work unsuccessfully for these data, I created a tool, VIOLA[39], to label imagery with my team. This additional data supported the development of SPOT[38], the first learning-based object detector for animals and humans in aerial, thermal videos. Our partners field tested SPOT in Africa, where it outperformed their prior tool. To build upon these results, I created AirSim-W, a simulated environment with a thermal model[37]. I then collected simulated imagery and sampled frames for the final balanced training set using a novel mixed integer linear program (MILP), as frames must remain grouped by video in either the train or test set and may have multiple objects, while the variety of videos represented in training should be maximized. A MILP-balanced dataset led to similar results as the full dataset, implying that a variety of frames may allow for less labeled data. Real and simulated data, and benchmarks, are available[40].

As another example, micronutrient deficiency (MND), or a lack of vitamins and minerals, is a significant public health concern affecting $>$ 2 billion people worldwide. Unfortunately, it typically requires blood draws to diagnose. A proxy used in past work is surveys, for example about food consumed, which is also expensive. To provide data to public health officials more scalably, our interdisciplinary team proposed a regional-level MND prediction system[36] based on another data modality: satellite data. This may at first seem unrelated, as MND status pertains to an individual's characteristics which cannot be viewed by satellite. However, public health literature shows that access to forests and markets, for example, may affect MND status, and these are visible by satellite. I therefore gathered satellite data and a limited amount of ground truth data from blood draws, collected by our public health partners at 300 unique locations in Madagascar. Given the limited amount of ground truth data, I first performed feature selection on the satellite data to prevent overfitting. I designed a feature selection algorithm based on k-medoids, which allowed us to choose one feature from multiple correlated features, aiding interpretability. MND prediction based on the selected features performed comparably to human expert-selected features, facilitating scalability. I also utilized domain adaptation with a shallow multi-layer perceptron, allowing us to use ground truth data from multiple ecological regions. MND prediction improved over a baseline of survey responses, which is promising for future interventions, and shows the promise of looking at additional data modalities and data augmentation generally.

0.2.4 Interdisciplinary, inclusive teams

Finally, together with an interdisciplinary team, I proposed the PACT framework to guide further work in AI for social impact, and especially to center communities' needs in AI systems like these[43]. It is based on the *capabilities approach*, which emphasizes the *capability* of people to pursue the lives they envision, and the participatory approach, in which community members are partners in the development and evaluation of AI systems. I believe this is imperative for future AI systems.

0.3 Thesis Outline

The thesis follows the structure of this summary and is organized as follows: Chapter 1 introduces modeling uncertainty in all components of these systems, particularly during multi-agent reasoning. Chapter 2 describes human-agent interaction, and Chapter 3-7 describe data augmentation to address uncertainty. Finally, Chapter 8 describes a general framework for developing and deploying AI systems with interdisciplinary, inclusive teams. Related work and background information is provided in each chapter. Chapter 9 contains a summary and future directions.

1

Uncertain Real-Time Information in Signaling Games

1.1 Introduction

Conservation drones have been deployed in South Africa to prevent wildlife poaching in national parks (Fig. 1.1). The drones, equipped with thermal infrared cameras, fly throughout the park at night when poaching typically occurs. Should anything suspicious be observed in the videos, nearby park rangers can prevent poaching, and a warning signal (e.g., drone lights) can be deployed for deterrence[5]. This requires a great deal of planning and coordination, as well as constant video monitoring. Rather than constant monitoring, we have worked with Air Shepherd to deploy an automatic detection system to locate humans and animals in these videos. Although an automatic detection system is helpful, its detections are uncertain. Potential false negative detections, in which the system fails to detect actual poachers, may lead to missed opportunities to deter or prevent poaching. This work is motivated by this real-world deployment of drones for conservation.

Security challenges similar to those in conservation must be addressed around the world, from protecting large public gatherings such as marathons[336] to protecting cities. Security game models have been shown to be effective in many of these real-world domains[286,52]. Recently, these models have begun to take into account real-time information, for example by using information from footprints when tracking poachers, or images from sensors[310,20]. In particular, signaling based on real-time information, e.g., signaling to indicate the presence of law enforcement[326], has been introduced and established as a fundamental area of work.

Despite the rising interest in real-time information and signaling, unfortunately, security games literature has failed to consider uncertainty in sensing real-time information and signaling, hindering real-world applicability of the game models. Previously, only some types of uncertainty have been considered, such as uncertainty in the adversary's observation of the defender's strategy, adversary's payoff values, or adversary's rationality[337,218,333]. However, there are fundamentally new insights when handling uncertainties w.r.t. real-time sensing and signaling, which we discuss at the end of this section.

We therefore focus on uncertainty in security games, in which real-time information comes from sensors that alert the defender when an adversary is detected and can also send warning signals to the adversary to deter the attack in real time. We consider both uncertainty in the sensor's detection of

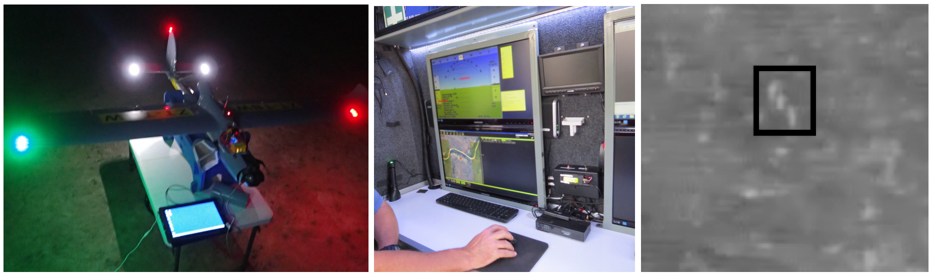


Figure 1.1: A drone and drone team member who are currently searching for poachers in a South African park at night.

adversaries (henceforth detection uncertainty) and uncertainty in the adversaries' observation of the sensor's signals (henceforth observational uncertainty), and show that ignoring uncertainty hurts the defender's expected utility. In our motivating domain of wildlife conservation with drones, automatic detection algorithms may make incorrect detections because humans in thermal infrared frames look similar to other objects (e.g., Fig. 1.1) and may even be occluded by other objects from the aerial perspective. The drone is also used to emit light to deter poachers, but such signals could sometimes be difficult for poachers to see in the wild, e.g., when trees block the sight.

We make contributions in (i) modeling, (ii) theoretical analysis, (iii) algorithmic design, and (iv) empirical evaluation. (i) We are the first to model uncertainty in sensing and signaling settings for security games. We introduce a novel reaction stage to the game model and construct a new signaling scheme, allowing the defender to mitigate the impact of uncertainty. In fact, this signaling scheme *exploits uncertain real-time information and the defender's informational advantage*. For example, both the defender and adversary may know that there is detection uncertainty; however, the defender has an informational advantage in knowing that she has or has not actually detected the adversary, which she can exploit via a signaling scheme to "mislead" the adversary who is uncertain as to whether he has been detected. (ii) We provide several theoretical results on the impact of uncertainties, e.g., the loss due to ignoring observational uncertainty can be arbitrarily large, illustrating the need to handle uncertainty. (iii) To compute the defender's optimal strategy given uncertainty, we develop a novel algorithm, GUARDSS, that not only uses six states to represent the type of protection a target has in a defender's pure strategy but also uses a new matching technique in a branch-and-bound framework. (iv) We conduct extensive experiments on simulation based on our real-world deployment of a conservation drone system[*].

1.2 Related Work

Among the rich literature of Stackelberg security games (SSGs)[286,52], SSGs with real-time information have been studied recently. Some recent work in deception for cybersecurity, such as[74,290], considers strategic signaling with boundedly rational adversaries and adversaries with different objectives and abilities, but no sensing is required to identify adversaries; rather, the systems may interact with both normal and adversarial users. Some other work relies on human patrollers for real-time information[343,310], and others rely on sensors that can notify the patroller when an opponent is detected[82,20,83]. Sensor placement[125] and drone patrolling[259] have also been studied. Spatial and detection uncertainties in alarms are examined in[21,22]. In all of these works, the sensors are only used to collect information, and do not actively and possibly deceptively disseminate information to the adversary. One work that does consider mobile sensors with detection and signaling capability is[326]. However, it does not consider uncertainty in detection, which limits its capability in real-world settings. We add a new reaction stage and signaling strategy without detection, and compactly encode the different states that the defender resources can have at a target. Our model is therefore strictly more general than that in[326].

Our work is also related to multistage game models, e.g., defender-adversary-defender sequential games (DAD)[49,7]. In DAD, the defender and adversary take turns to commit to strategies, while in our game, the defender commits to a strategy of all stages at once. Extensive-form games (EFGs)

[*] https://github.com/exb7900/guardss-aaai2020

also naturally model sequential games [164,51,207], and algorithms exist to efficiently solve the Stackelberg equilibrium in general two-player EFGs [59,58]. However, GUARDSS is more scalable than the general EFG approach in this case (see Appendix). Finally, past work relied upon online, recursive cognitive modeling [288,285], whereas our calculations are done offline and based largely on utilities.

1.3 Model

We consider a security game played between a defender and an adversary who seeks to attack one target. The defender has k human patrollers and l sensors to be allocated to targets in set $[N] = \{1, 2, ..., N\}$. The sensor is the same as a drone in our motivation domain, and the adversary is the same as a poacher. Let $U^{d/a}_{+/-}(i)$ be the defender/adversary (d/a) utility when the defender successfully protects/fails to protect ($+/-$) the attacked target i. By convention, we assume $U^d_+(i) \geq 0 > U^d_-(i)$ and $U^a_+(i) \leq 0 < U^a_-(i)$ for any $i \in [N]$. The underlying geographic structure of targets is captured by an undirected graph $G = (V, E)$ (e.g., Fig. 1.4). A patroller can move to any neighboring target and successfully interdict an attack at the target at no cost.

Sensors cannot interdict an attack, but they can notify nearby patrollers to respond and signal to deter the adversary. If the adversary is deterred by a signal (e.g., runs away), both players get utility 0. In practice, often one signal (σ_1, e.g., illuminating the lights on the drone) is a warning that a patroller is nearby, while another signal (σ_0, e.g., turning no lights on) indicates no patroller is nearby, although these may be used deceptively. Theoretically, [145] also showed two signals suffice (without uncertainty). We thus use two signals: σ_1 is a *strong signal* and σ_0 is a *weak signal*. When the adversary chooses one target to attack, he encounters one of four *signaling states*, based on the target either having a patroller, nothing, or a drone. The adversary may encounter: (1) a patroller and immediately get caught (state p); (2) nothing (state n); (3) a drone with signal σ_0 (state σ_0); (4) a drone with signal σ_1 (state σ_1). The adversary is caught immediately at state p, so there is no signal.

Therefore, we omit p and let $\Omega = \{\mathrm{n}, \sigma_0, \sigma_1\}$ be the set of signaling states.

1.3.1 Modeling Uncertainty

In this paper, we focus on two prominent uncertainties motivated directly by the use of conservation drones. The first is the *detection uncertainty*, when there is a limitation in the sensor's capability, e.g., a detection could be incorrect due to the inaccuracy of image detection techniques in the conservation domain[40,38,227]. We consider only false negative detection in this paper because patrollers often have access to sensor videos, so the problem of false positives can be partly resolved with a human in the loop. In contrast, verifying false negatives is harder, e.g., the adversary is easy to miss in the frame (Fig. 1) or is occluded. We therefore denote the false negative rate as γ for any sensor[†].

The second type of uncertainty we consider is the *observational uncertainty*, where the true signaling state of the target may differ from the adversary's observation (e.g., a poacher may not be able to detect the drone's signal). We use $\hat{\omega}$ to denote the adversary's observed signaling state, and use ω to denote the true signaling state based on the defender signaling scheme. We introduce uncertainty matrix Π to capture observational uncertainty. The uncertainty matrix Π will contain the conditional probability $\Pr[\hat{\omega}|\omega]$ for all $\hat{\omega}, \omega \in \Omega$ to describe how likely the adversary will observe a signaling state $\hat{\omega}$ given the true signaling state is ω.

$$\Pi = \begin{bmatrix} \Pr[\hat{\omega} = \mathrm{n}|\mathrm{n}] & \Pr[\hat{\omega} = \mathrm{n}|\sigma_0] & \Pr[\hat{\omega} = \mathrm{n}|\sigma_1] \\ \Pr[\hat{\omega} = \sigma_0|\mathrm{n}] & \Pr[\hat{\omega} = \sigma_0|\sigma_0] & \Pr[\hat{\omega} = \sigma_0|\sigma_1] \\ \Pr[\hat{\omega} = \sigma_1|\mathrm{n}] & \Pr[\hat{\omega} = \sigma_1|\sigma_0] & \Pr[\hat{\omega} = \sigma_1|\sigma_1] \end{bmatrix}$$

Considering an arbitrary uncertainty matrix may unnecessarily complicate the problem, since some uncertainties never happen. We thus focus on a restricted class of uncertainty matrices that

[†] False negative rate: P(no detection | poacher is present).

are natural in our domain.[‡] In our uncertainty model, we assume that a weak signal will never be observed as strong; moreover, n (the signaling state without any resource) will never be observed as strong or weak. As a result, the uncertainty matrix Π can be reduced to the following form, parameterized by κ, λ, μ, where $\kappa = \Pr[\hat{\omega} = \mathrm{n}|\sigma_0]$, $\lambda = \Pr[\hat{\omega} = \mathrm{n}|\sigma_1]$, $\mu = \Pr[\hat{\omega} = \sigma_0|\sigma_1]$:

$$\Pi_{\kappa\lambda\mu} = \begin{bmatrix} 1 & \kappa & \lambda \\ 0 & 1-\kappa & \mu \\ 0 & 0 & 1-\lambda-\mu \end{bmatrix}$$

As a result of this uncertainty, the adversary may not behave as expected. For example, if he knows that he has difficulty seeing the strong signal, he may decide to attack only when there is no drone, whereas typically we would expect him to attack on a weak signal. Therefore, let $\boldsymbol{\eta} \in \{0,1\}^3$ be the vector that depicts adversary behavior for each observation $\{\mathrm{n}, \sigma_0, \sigma_1\} \in \Omega$, where 1 represents attacking, and 0 represents running away. So, $\boldsymbol{\eta} = \mathbf{1}$ means an adversary will attack no matter what signaling state is observed, and $\boldsymbol{\eta} = \mathbf{0}$ means an adversary will never attack.

1.3.2 Reaction Stage

Uncertainty motivates us to add an explicit reaction stage during which the defender can respond *or* re-allocate patrollers to check on extremely uncertain sensors or previously unprotected targets, for example. The timing of the game is summarized in Fig. 1.2. In words, (i) the defender commits to a mixed strategy and then executes a pure strategy allocation; (ii) the adversary chooses a target to attack; (iii) the sensors detect the adversary with detection uncertainty; (iv) the sensors signal based on the signaling scheme; (v) *the defender re-allocates patrollers based on sensor detections and matching;* (vi) the adversary observes the signal with observational uncertainty; (vii) the adversary chooses to ei-

[‡] Most results can be extended to general uncertainty matrices.

ther continue the attack or run away. In (v), if a sensor detects the adversary, then nearby patroller(s) (if any) always go to that target, and the game ends; *or if no sensors or patrollers detect the adversary, the patroller moves to another target to check for the adversary.* The adversary reaction occurs after the defender reaction because the adversary reaction does not affect the defender reaction in the current model. In other words, there is no cost in reallocating the defender even if the adversary runs away, so the defender should begin moving right away.

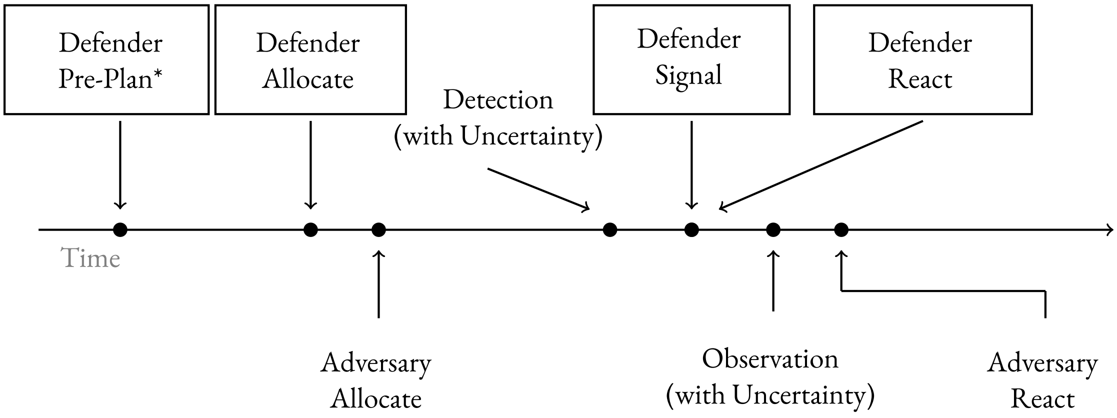


Figure 1.2: Game timing. Top and bottom are defender and adversary actions, respectively. *Defender fixes strategy offline.

1.3.3 Defender and Adversary Strategies

Defender Strategy: The strategy space consists of randomized resource allocation and re-allocation, and signaling. A deterministic resource allocation and re-allocation strategy (henceforth, a *defender pure strategy*) consists of allocating the patrollers to k targets, the sensors to l targets, and the neighboring target to which each patroller moves if no adversaries are observed. Re-allocation can be equivalently thought of as matching each patroller's original target to a neighboring target. A patroller goes to the matched target only if the adversary is not observed, and may respond to any nearby sensor detection, regardless of matching.

As a result of this rich structure, a pure strategy in the model needs to represent not only if the target is assigned a patroller (p), nothing (n), or a sensor (s), but also the allocation in neighboring targets. We compactly encode this pure strategy via 6 possible *allocation states* for each target. Let $\Theta = \{\mathrm{p}, \mathrm{n}+, \mathrm{n}-, \bar{\mathrm{s}}, \mathrm{s}+, \mathrm{s}-\}$ denote the set of all possible allocation states of an individual target. The target is assigned a patroller (p), nothing (n), or a sensor (s). If there is no patroller near a sensor ($\bar{\mathrm{s}}$), then no one can respond to the sensor's detection. If there is a nearby patroller, the target is either matched ($\mathrm{n}+, \mathrm{s}+$) or not matched ($\mathrm{n}-, \mathrm{s}-$). Therefore, each target is in one of the allocation states in Table 1.1. For example, n+ is the state of a target which was not allocated a patroller or sensor, but in the reaction stage has a patroller from a neighboring target ("patroller matched").

	Covered By:	Near Patroller?	Patroller Matched?	Protected Overall?
p	Patroller	N/A	N/A	Yes
n+	Nothing	Yes	Yes	Yes
n-	Nothing	N/A	No	No
$\bar{\mathrm{s}}$	Sensor	No	N/A	No
s-	Sensor	Yes	No	Yes*
s+	Sensor	Yes	Yes	Yes

Table 1.1: Allocation State, *protected if sensor detects

Given Θ, a defender pure strategy can be compactly represented with an allocation state vector $\mathbf{e} \in \Theta^N$, in which $e_i \in \Theta$ denote the allocation state of a target $i \in [N]$. Let $\mathcal{E} \subseteq \Theta^N$ be the set of feasible allocation state vectors that corresponds to defender pure strategies. Note that not all vectors in Θ^N correspond to a feasible defender strategy due to the limited number of patrollers and sensors. A *defender mixed strategy* is thus a distribution over $\mathcal{E}$ and can be described by $\{q_e\}_{\mathbf{e}\in\mathcal{E}}$ where $q_{\mathbf{e}}$ is the probability of playing pure strategy $\mathbf{e} \in \mathcal{E}$. Similarly, a defender mixed strategy can also be compactly represented by a marginal probability vector x, where x_i^θ represents the marginal probability that target i is in the allocation state $\theta \in \Theta$. This is similar to the coverage vector used in basic SSGs with schedules[138]. We introduce the constraints that x needs to satisfy to be a valid mixed strategy in Section 1.5.

The defender also deploys a signaling process w.r.t. each target i. The defender's signaling strategy can be specified by probabilities ψ_i^{s-}, ψ_i^{s+}, and $\psi_i^{\bar{s}}$. ψ_i^{s-} is the *joint* probability of allocation state s$-$ and sending signal σ_0 together conditioned on the sensor detecting an adversary, i.e., $\Pr[s- \wedge \sigma_0|\text{detected}]$. To be a valid signaling strategy, $\psi_i^{s-} \in [0, x_i^{s-}]$. Note that $x_i^{s-} - \psi_i^{s-}$ will be the joint probability of realized state s$-$ and sending signal σ_1, together conditioned on detection. The conditional probability of sending σ_0 given the target is in state s$-$ and it is detected is ψ_i^{s-}/x_i^{s-}. We use the joint probability instead of the conditional probability as it results in linear terms for the optimal defender strategy. *Because of detection uncertainty, we add the option to signal without detecting the adversary.* Let $\phi_i^{\theta} \in [0, x_i^{\theta}]$ be the joint probability of allocation state θ and sending signal σ_0 conditioned on the sensor not detecting an adversary, for all $\theta \in \{\bar{s}, s-, s+\}$. We use χ to denote the allocation, reaction, and signaling scheme, or *defender's deployment strategy*: $\chi = (x, \psi, \phi)$.

Adversary Strategy: Recall the adversary has the allocation and reaction stages. In the allocation stage, the adversary chooses a target to attack based on the defender deployment strategy χ. He will be caught if the target is at state p. When the adversary is not caught, he may observe any of the signaling states $\hat{\omega} \in \Omega$. Based on his observation, the adversary then has a choice in the reaction stage to run away or continue the attack. The adversary knows the defender mixed strategy χ when choosing a target to attack, and he can observe the realization of the target (with uncertainty) when choosing to attack or run away. Since this is a Stackelberg game and the defender commits to allocation and signaling schemes, it suffices to consider only the adversary's pure responses.

1.4 Why Do We Need to Handle Uncertainty

In this section, we prove several theoretical properties regarding how uncertainties affect the defender's optimal strategy and utility. All formal proofs are deferred to the Appendix. Let $\chi^*(\gamma, \Pi)$

be the optimal allocation under detection uncertainty of γ and observational uncertainty Π. Let $\mathsf{DefEU}(\chi, \gamma, \Pi)$ be the defender expected utility when the actual uncertainties are γ, Π and the defender's deployment is χ. Let $\Pi_0 = \mathbf{I}$ denote no observational uncertainty. We assume in Propositions 1 and 2 and Theorem 1 that $\Pi = \Pi_0$ and analyze detection uncertainty, so omit for conciseness. We first show the loss due to ignoring detection uncertainty.

Proposition 1. *Let $\chi_0^* = \chi^*(0)$ be the defender optimal deployment when no uncertainties exist. There exist instances where $\mathsf{DefEU}(\chi_0^*, \gamma) < \mathsf{DefEU}(\chi^*(\gamma), \gamma)$ for some γ.*

In fact, $\mathsf{DefEU}(\chi^*(\gamma), \gamma) - \mathsf{DefEU}(\chi_0^*, \gamma) \geq \gamma \cdot \max_{i \in [N]} |U^d_-(i)|$ for some instance. If we ignore γ, we do not signal when we do not detect an adversary. Furthermore, the defender would never match a patroller to a target with a sensor (s+) in χ_0^*. Thus, if we ignore uncertainty, there can be a steep penalty; in contrast, with the optimal strategy considering uncertainty, if the false negative rate is high, we may match a patroller to a target to confirm the presence of an adversary. Given the adversary's knowledge of the defender mixed strategy, the adversary is therefore more likely to run away.

Our next result (Theorem 1) shows that the defender expected utility is non-increasing as detection uncertainty γ increases. As a byproduct of the proof for Theorem 1, we also show that the optimal solution may change as detection uncertainty changes. This illustrates the necessity of an algorithm for dealing with detection uncertainties.

Theorem 1. $\mathsf{DefEU}(\chi^*(\gamma), \gamma) \geq \mathsf{DefEU}(\chi^*(\gamma'), \gamma')$ *for any $\gamma' > \gamma$ in any problem instance.*

Proposition 2. *$\chi^*(\gamma)$ differs from $\chi^*(\gamma')$ for any $\gamma' > \gamma$ when x_t^{s-} is nonzero for $\chi^*(\gamma')$, where target t is the adversary best responding target in $\chi^*(\gamma')$.*

The intuition underlying the proof of Theorem 1 is that if we have a drone with a low false negative rate, then we can simulate a drone with a high false negative rate by ignoring some of its detections.

The optimal solution for drones with a low false negative rate cannot be worse than that for drones with a high false negative rate.

We now show several results for observational uncertainty. First, we show that the loss due to observational uncertainty can be arbitrarily large.

Proposition 3. *There exists* Π *such that the loss due to ignoring observational uncertainty is arbitrarily large. In other words,* $\mathsf{DefEU}(\chi^*(\gamma_0, \Pi), \gamma_0, \Pi)$ - $\mathsf{DefEU}(\chi^*(\gamma_0, \Pi_0), \gamma_0, \Pi) > M$*,* $\forall M > 0$.

The original signaling strategy tries to ensure the adversary only attacks when he observes the weak signal, σ_0, or nothing, n. However, with observational uncertainty, this may not be true because the true signal may be σ_1, but the adversary may have observed it mistakenly as σ_0. Therefore, we need to enforce different adversary behaviors in order to obtain a better solution quality.

Now, we examine the adversary's behavior given a fixed deployment χ as observational uncertainty changes. Let (t, η) represent an adversary strategy of attacking target t and behaving according to η. Theorems 2 and 3 show that if we do not consider observational uncertainty, then the adversary behavior is more likely to converge to always attacking ($\eta = 1$) as observational uncertainty increases, where higher observational uncertainty means the adversary cannot distinguish between signaling states. Theorems 2 and 3 show that a deployment χ that does not consider observational uncertainty is more likely to result in this worst-case behavior of $\eta = 1$.

Theorem 2. *For any fixed deployment* χ*, if the adversary's best response is* $(t, 0)$ *or* $(t, 1)$ *at the Stackelberg equilibrium with* Π_0*, then it stays as an equilibrium for any* Π'*.*

Note that $\eta = 0$ and $\eta = 1$ result in an action that is independent of the adversary's observation. Thus, no matter what the adversary observes, the adversary can obtain the same utility with $\eta = 0$ or $\eta = 1$. It's only left to show that the adversary cannot get strictly better utility in Π' with a different adversary behavior. Intuitively, Π_0 implies a perfect observation, thus the adversary cannot get better utility than the perfect observation. So, if $(t, 1)$ or $(t, 0)$ is a Stackelberg equilibrium, the defender can safely deploy the same strategy for any uncertainty matrix Π', without any loss in her expected utility.

Even if $(t, 0)$ or $(t, 1)$ is not a best response with Π_0, $(t, 1)$ may still be a best response at high levels of uncertainty. First, we say a target t is a *weak-signal-attack target* if $\mathsf{AttEU}(\sigma_0) \geq 0$ at t. Note that if $\mathsf{AttEU}(\sigma_0) \geq 0$, then the adversary will either always attack at $\hat{\omega} = \sigma_0$, or is indifferent between attacking and running away. We say χ is a *weak-signal-attack deployment* if all targets are weak-signal-attack targets.

Theorem 3. *If $(t, 1)$ is a best response for $\Pi_{\kappa\lambda\mu}$ and χ is a weak-signal-attack deployment, then $(t, 1)$ is a best response for $\Pi_{\kappa'\lambda'\mu'}$ and χ for all $\kappa' \geq \kappa$, $\lambda' \geq \lambda$, $\mu' \geq \mu$.*

In our model of observational uncertainty, more uncertainty means that the adversary sees a weak signal more often. Further, the adversary always attacks when he observes a weak signal. Thus, if the adversary is always attacking with less uncertainty, he will only attack more often with more uncertainty. However, in order to obtain predictable adversary behavior, we need to show that a weak-signal-attack deployment always exists as an optimal solution. In other words, Theorem 3 holds if there is weak-signal-attack deployment, so we now have to show that such a deployment exists.

Proposition 4. *There always exists an optimal solution that is a weak-signal-attack deployment with Π_0.*

The intuition behind the proof is that we can always decrease the probability of a weak signal such that we either do not send a weak signal, or the adversary attacks when he observes a weak signal. This holds optimally because when observational uncertainty is Π_0, signals are interchangeable. To summarize, if the adversary behavior is 0 or 1, then the adversary behavior is independent of observational uncertainty. We may see this behavior emerge as uncertainty increases.

1.5 How to Handle Uncertainty

We provide a solution approach based on the well-known multiple LPs approach from[72]. In particular, for each target $t \in [N]$, we compute the optimal defender strategy given that the adversary's best response is t. Then, the optimal defender strategy is the mixed strategy that leads to the maximum defender expected utility among all $t \in [N]$. The problem is NP-hard without uncertainty[326], thus our ultimate goal is to develop an efficient algorithm to solve the problem. For expository purposes, we first focus on presenting the LP for detection uncertainty.

1.5.1 Detection Uncertainty

Using notation from Section 1.3.3, we first formulate each player's utility function by breaking it into three parts according to signaling states: 1) no sensor is allocated (states $\mathrm{n}(+/-)$ and p, which we denote by $-s$); 2) sensor is allocated and sends σ_0; and 3) sensor is allocated and sends σ_1.

1. $U^{d/a}_{\text{-}s}(i) = x^{\mathrm{p}}_i \cdot U^{d/a}_{+}(i) + x^{\mathrm{n}+}_i \cdot U^{d/a}_{+}(i) + x^{\mathrm{n}-}_i \cdot U^{d/a}_{-}(i)$ is the expected defender/adversary utility of target i being attacked over states when i has no sensor (p, n+, n−).

2. $U^{d/a}_{\sigma_0}(i) = (1-\gamma) \cdot [\psi^{\mathrm{s}+}_i \cdot U^{d/a}_{+}(i) + \psi^{\mathrm{s}-}_i \cdot U^{d/a}_{+}(i) + \psi^{\bar{\mathrm{s}}}_i \cdot U^{d/a}_{-}(i)] + \gamma \cdot [\phi^{\mathrm{s}+}_i \cdot U^{d/a}_{+}(i) + \phi^{\mathrm{s}-}_i \cdot U^{d/a}_{-}(i) + \phi^{\bar{\mathrm{s}}}_i \cdot U^{d/a}_{-}(i)]$ is the defender/adversary expected utility when the adversary attacks target i and the defender signals σ_0.

3. $U^{d/a}_{\sigma_1}(i) = (1-\gamma) \cdot [(x^{\mathrm{s}+}_i - \psi^{\mathrm{s}+}_i) \cdot U^{d/a}_{+}(i) + (x^{\mathrm{s}-}_i - \psi^{\mathrm{s}-}_i) \cdot U^{d/a}_{+}(i) + (x^{\bar{\mathrm{s}}}_i - \psi^{\bar{\mathrm{s}}}_i) \cdot U^{d/a}_{-}(i)] + \gamma \cdot [(x^{\mathrm{s}+}_i - \phi^{\mathrm{s}+}_i) \cdot U^{d/a}_{+}(i) + (x^{\mathrm{s}-}_i - \phi^{\mathrm{s}-}_i) \cdot U^{d/a}_{-}(i) + (x^{\bar{\mathrm{s}}}_i - \phi^{\bar{\mathrm{s}}}_i) \cdot U^{d/a}_{-}(i)]$

In words, 2) and 3) are the sum of expected utility on a detection and the sum of expected utility on no detection. In 3), in the no detection case, the defender exploits information asymmetry in signaling σ_1. In particular, the defender knows that there is no detection, but in sending σ_1 to indicate a detection, relies on the uncertainty the adversary faces in determining if there was a detection. We are now ready to describe an (exponentially-large) linear program (LP) formulation for computing the optimal defender strategy assuming best adversary response t (not (t, η) since only detection uncertainty):

$$\max_{x,q,\psi,\phi} U^d_{\text{-}s}(t) + U^d_{\sigma_0}(t) \tag{1.1}$$

$$\text{s.t.} \quad \textstyle\sum_{\mathbf{e}\in\mathcal{E}:e_i=\theta} q_{\mathbf{e}} = x^\theta_i \qquad \forall\, \theta \in \Theta, \forall\, i \in [N] \tag{1.2}$$

$$\textstyle\sum_{\mathbf{e}\in\mathcal{E}} q_{\mathbf{e}} = 1 \tag{1.3}$$

$$q_{\mathbf{e}} \geq 0 \qquad \forall\, \mathbf{e} \in \mathcal{E} \tag{1.4}$$

$$U^a_{\sigma_0}(i) \geq 0 \qquad \forall i \neq t \tag{1.5}$$

$$U^a_{\sigma_1}(i) \leq 0 \qquad \forall i \neq t \tag{1.6}$$

$$U^a_{\text{-}s}(t) + U^a_{\sigma_0}(t) \geq U^a_{\text{-}s}(i) + U^a_{\sigma_0}(i) \quad \forall i \neq t \tag{1.7}$$

$$0 \leq \psi^\theta_i \leq x^\theta_i \quad \forall\, \theta \in \{\bar{\mathrm{s}}, \mathrm{s}-, \mathrm{s}+\}, \forall\, i \in [N] \tag{1.8}$$

$$0 \leq \phi^\theta_i \leq x^\theta_i \quad \forall\, \theta \in \{\bar{\mathrm{s}}, \mathrm{s}-, \mathrm{s}+\}, \forall\, i \in [N] \tag{1.9}$$

The objective function (1.1) maximizes defender expected utility. Since the adversary is running away when he observes σ_1, $U^d_{\sigma_1} = 0$. Constraints (1.2)-(1.4) enforce that the randomized resource allocation is feasible ($\mathcal{E}$ has exponential number of elements); (1.5)-(1.6) guarantee that σ_1, σ_0 result in the adversary best responses of running away and attacking[§]; (1.7) ensures the adversary expected utility at target t is bigger than at any other target i, thus t is adversary's best response; (1.8)-(1.9) ensure a feasible signaling scheme.

[§] Although we minimize this behavior, we still model it.

1.5.2 Acceleration via Branch and Price

We now describe the branch-and-price solution framework, which can be used for both uncertainty scenarios. There are two main challenges in efficiently solving the LP (1.1)-(1.9). First, the total number of possible e is $O(6^N)$. Second, we will need to solve N LPs (for each $t \in [N]$). Solving many of these large LPs is a significant barrier for scaling up. We therefore introduce Games with Uncertainty And Response to Detection with Signaling Solver (GUARDSS), which employs the branch-and-price framework. This framework is well-known for solving large-scale optimization programs, but the main challenges of applying this framework are to (1) design the subroutine called the *secondary problem*[¶], and to (2) carefully design an upper bound for pruning LPs.

First, for one LP w.r.t. a specific t, to address the issue of the exponential size of set $\mathcal{E}$, we adopt the column generation technique. At a high level, we start by solving the LP for a small subset $\mathcal{E}' \subset \mathcal{E}$, and then search for a pure strategy $e \in \mathcal{E} \setminus \mathcal{E}'$ such that adding e to $\mathcal{E}'$ improves the optimal objective value strictly. This procedure continues until convergence, i.e., no objective value improvement. The key component in this technique is an algorithm to search for the new pure strategy, which is a specially-crafted *secondary problem* derived from LP duality.

Secondary Problem: Given weights $\alpha_i^\theta \in \mathbb{R}$ for $\theta \in \Theta$, for each target i, solve the *weight maximization problem*:

$$\max_{e \in \mathcal{E}} \sum_{\theta \in \Theta} \sum_{i : e_i = \theta} \alpha_i^\theta \tag{1.10}$$

Note that $\{\alpha_i^\theta\}_{\theta \in \Theta}$ are the optimal dual variables for the previous LP constraint (1.2). We want to solve this without enumerating all of the elements in $\mathcal{E}$. Despite the added complexity compared to classic SSGs, in this section, we compactly represent this secondary problem as a mixed integer

[¶] Updated from historical terminology.

linear program (MILP). To formulate the MILP, we introduce six binary vectors $\mathbf{v}^{\mathrm{p}}$, $\mathbf{v}^{\mathrm{n}+}$, $\mathbf{v}^{\mathrm{n}-}$, $\mathbf{v}^{\bar{\mathrm{s}}}$, $\mathbf{v}^{\mathrm{s}-}, \mathbf{v}^{\mathrm{s}+} \in \{0,1\}^N$ to encode for each target whether it is in each allocation state. For example, target i is at allocation state $\bar{\mathrm{s}}$ if and only if $v_i^{\bar{\mathrm{s}}} = 1$. The main challenge then is to properly set up linear (in)equalities of these vectors to precisely capture their constraints and relations. The capacity for each resource type results in two constraints (number of patrollers and sensors):

$$\textstyle\sum_{i\in[N]} v_i^{\mathrm{p}} \leq k \tag{1.11}$$

$$\textstyle\sum_{i\in[N]}(v_i^{\bar{\mathrm{s}}} + v_i^{\mathrm{s}-} + v_i^{\mathrm{s}+}) \leq l \tag{1.12}$$

Moreover, each target must be at one of these states:

$$v_i^{\mathrm{p}} + v_i^{\mathrm{n}-} + v_i^{\mathrm{n}+} + v_i^{\bar{\mathrm{s}}} + v_i^{\mathrm{s}-} + v_i^{\mathrm{s}+} = 1 \quad \forall i \in [N] \tag{1.13}$$

Due to the reaction stage, we have to add constraints to specify (a) which targets have a patroller at a neighboring target; (b) which patroller goes to which nearby target if both sensors and patrollers do not detect the adversary. For (a), the non-zero entries of $A \cdot \mathbf{v}^{\mathrm{p}}$ specify the targets with a patroller nearby, where A is the adjacency matrix of the underlying graph. Since three vectors encode the states requiring a nearby patroller, we have this constraint:

$$A \cdot \mathbf{v}^{\mathrm{p}} \geq \mathbf{v}^{\mathrm{n}+} + \mathbf{v}^{\mathrm{s}-} + \mathbf{v}^{\mathrm{s}+} \tag{1.14}$$

We ensure that a vertex with a patroller nearby cannot be $\mathbf{v}^{\bar{\mathrm{s}}}$:

$$A \cdot \mathbf{v}^{\mathrm{p}} \leq \mathbf{v}^{\mathrm{p}} + \mathbf{v}^{\mathrm{n}+} + \mathbf{v}^{\mathrm{n}-} + \mathbf{v}^{\mathrm{s}-} + \mathbf{v}^{\mathrm{s}+} \tag{1.15}$$

Constraint (b) means that patrollers must be "re-matched" to new vertices in the reaction stage. Specifically, targets in states p, n+, s+ must form a matching. To enforce this constraint, let G' be the directed version of G, i.e. for all $(i,j) \in E$ we have $(i,j), (j,i) \in E'$. *We further introduce edge variables $y_{(i,j)} \in \{0,1\}$ indicating whether the directed edge (i,j) is in the matching or not.* The matching constraint can be expressed by the following linear constraints[‖]

$$\textstyle\sum_{(i,j)\in E':j\in[N]} y_{(i,j)} = v_i^{\mathrm{p}} \qquad \forall i \in [N] \tag{1.16}$$

$$v_j^{\mathrm{n+}} + v_j^{\mathrm{s+}} \geq y_{(i,j)} \qquad \forall (i,j) \in E' \tag{1.17}$$

The resulting MILP for the secondary problem is as follows.

$$\begin{aligned}
\max_{\mathbf{v},y} \quad & \textstyle\sum_\theta \sum_i v_i^\theta \alpha_i^\theta & & \\
\text{s.t.} \quad & (1.11) - (1.17) & & \qquad (1.18) \\
& \mathbf{v}^\theta \in \{0,1\}^N & \forall \theta \in \Theta & \qquad (1.19) \\
& y_{(i,j)} \in \{0,1\} & \forall (i,j) \in E' & \qquad (1.20)
\end{aligned}$$

Second, to avoid solving LPs for all different targets $t \in [N]$, we use the branch and bound technique which finds an upper bound for each LP for pruning. The natural approach for finding an upper bound is to solve a relaxed LP corresponding to the original LP — in our case, essentially relax the original LP into its marginal space. As the set $\mathcal{E}$ is exponentially large, we relax variables and constraints corresponding to $\mathcal{E}$ in our LP. Concretely, we relax (1.2) - (1.4) into a polynomial number of variables and constraints. These variables and constraints are (1.18) - (1.20) with $\mathbf{v}^\theta$ replaced by $\mathbf{x}^\theta$. We first use the relaxed LP to efficiently compute an upper bound for each LP. After solving

[‖] Originally omitted (16.5): $\sum_{(i,j)\in E':i\in[N]} y_{(i,j)} \geq v_j^{\mathrm{n+}} + v_j^{\mathrm{s+}}$

each relaxed LP exactly, we solve original LPs chosen according to some heuristic order (typically the descending order of the relaxed optimal objective) using the column generation techniques, and we can safely prune out those LPs whose optimal relaxed value is less than the current largest achievable objective value. This process continues until no LP is left to solve, in which case the current largest objective value is optimal.

1.5.3 Detection and Observational Uncertainty

Finally, we briefly discuss the case with both uncertainties, as it can be solved in a similar way. Constraints (1.2)-(1.4) and (1.8)-(1.9) are the same. However, the remaining constraints must now account for adversary behavior, $\boldsymbol{\eta}$. For example, the utility functions $U^{d/a}_{\sigma_0}$ and $U^{d/a}_{\sigma_1}$ must change to incorporate adversary behaviors, and the objective function becomes that in (1.21) since the adversary may not run away when he observes σ_1 in the presence of observational uncertainty. Also, we add a constraint to ensure the adversary utilities are aligned with the adversary behavior $\boldsymbol{\eta} \in \{0, 1\}^3$. These are primarily notational changes. We therefore provide the full LP for this case in the Appendix.

$$\max_{x,q,\psi,\phi} \quad U^d_{-s}(t) + U^d_{\sigma_1}(t) + U^d_{\sigma_0}(t) \tag{1.21}$$

1.6 Experiments

We generate random Watts-Strogatz graphs, which have small-world properties to describe more complex environments, such as roads connecting far-away nodes. For all tests, we average over 20 random graphs and include p-values. Utilities are randomly generated with a maximum absolute value of 1090 and based on the idea that the losses from undetected attacks are higher than the utility of catching adversaries (similar to[325]). This is realistic to the situation of preventing poaching, as animals are worth more for ecotourism than for sale on the black market as discussed in the Appendix. Additionally, we see that if we test on a set of utilities that is slightly different from the original input, the defender's utility does not vary greatly. Fig. 1.3a-1.3b show timing tests run on a cluster with Intel(R) Xeon(R) CPU E5-2683 v4 @ 2.1 GHz with at most 16 GB RAM. We set the number of patrollers to be $k = \sqrt{N/2}$ and the number of drones to be $l = 2N/3 - k$. As shown, the full LP scales up to graphs of $N = 14$ only and exceeds the cutoff time limit of 3600s for all $N = 16$ graphs. Branch and price scales up to $N = 80$ and runs out of time for larger games, and a warm-up enhancement that greedily select an initial set of $\mathcal{E}$ further improves scalability and solves 13/40 graphs within cutoff time at $N = 90$ and $N = 100$. This is sufficient for middle-scale real-world problems, with further scalability being an interesting direction for future work. The heuristics provide the same solution as the full LP in most of the instances tested.

Next, we show the loss due to ignoring uncertainty empirically. In Figs. 1.3c-1.3d we compare $\mathsf{DefEU}(\chi^*(\gamma, \Pi), \gamma, \Pi)$ computed by GUARDSS and $\mathsf{DefEU}(\chi^*(0, \Pi_0), \gamma, \Pi)$, the defender expected utility when ignoring uncertainty for graphs with $N = 10, k = 1, l = 3$. We consider only one type of uncertainty at a time (e.g., $\gamma = 0$ when varying observational uncertainty). For detection uncertainty, GUARDSS's defender expected utility only decreases by 12%, whereas ignoring uncertainty decreases by 210% when γ varies from 0 to 0.9 (p $\leq 1.421\text{e}{-03}$ for $\gamma \geq 0.2$ in Fig. 1.3c)**. Some initial analysis shows that it is robust in most of the cases when we slightly under- or overestimate γ (e.g., the differences in defender expected utility are typically within 5-6% when the estimate of gamma is off by 0.1 or 0.2), but further investigation on dealing with such uncertainty over uncertainty would be an interesting direction for future work. For observational uncertainty, GUARDSS's defender expected utility only decreases by 1%, whereas ignoring uncertainty decreases by 18% as the observational uncertainty, parameterized by κ ($\lambda = \frac{\kappa}{2}$, and $\mu = \frac{\kappa}{2}$) varies from 0 to 0.9 (p ≤ 0.058 for $\kappa \geq 0.4$ in Fig. 1.3d).

**% change once normalized by largest defender/adversary utility.

We also observe that when ignoring detection uncertainty, the adversary's best response is typically a target with a sensor, which implies that the adversary is taking advantage of the defender's ignorance of uncertainty. In fact, there is a statistically significant ($p = 1.52e{-}08$) difference in the mean probability of a sensor at the adversary's best response when ignoring uncertainty (0.68) versus GUARDSS (0.19).

How does the defender avoid these challenges and achieve such a small performance drop with GUARDSS when facing uncertainty? Statistics of the resulting defender strategy as well as Fig. 1.3e indicate that the defender *exploits the uncertain real-time information and the information asymmetry*, including (a) frequently but not always sending patrollers to check important targets when there is no detection; (b) sending strong signals more frequently than the probability that the patroller will visit the target (either due to response to detection or planned reallocation in the case of no detection), leveraging the informational advantage in which the adversary does not know whether he is detected or whether a patroller is matched; (c) using different signaling schemes with and without detection, leveraging the information advantage that the adversary does not know whether he is detected. In the GUARDSS strategies in Figs. 1.3c-1.3d, the mean probability of the adversary's best response target being at state $s-$ (with sensor but without a matched patroller) is 0.04, versus 0.43 when ignoring uncertainty ($p = 2.70e{-}09$), indicating point (a). If we call the strong signal sent when there is no detection a *fake signal*, Fig. 1.3e shows that the probability of the strong signal an adversary observes is a fake signal is non-zero and increases in a non-linear fashion, indicating points (b) and (c). Also, note that the strong signal is used with nonzero probability on average on targets with a nonzero probability of having a drone present.

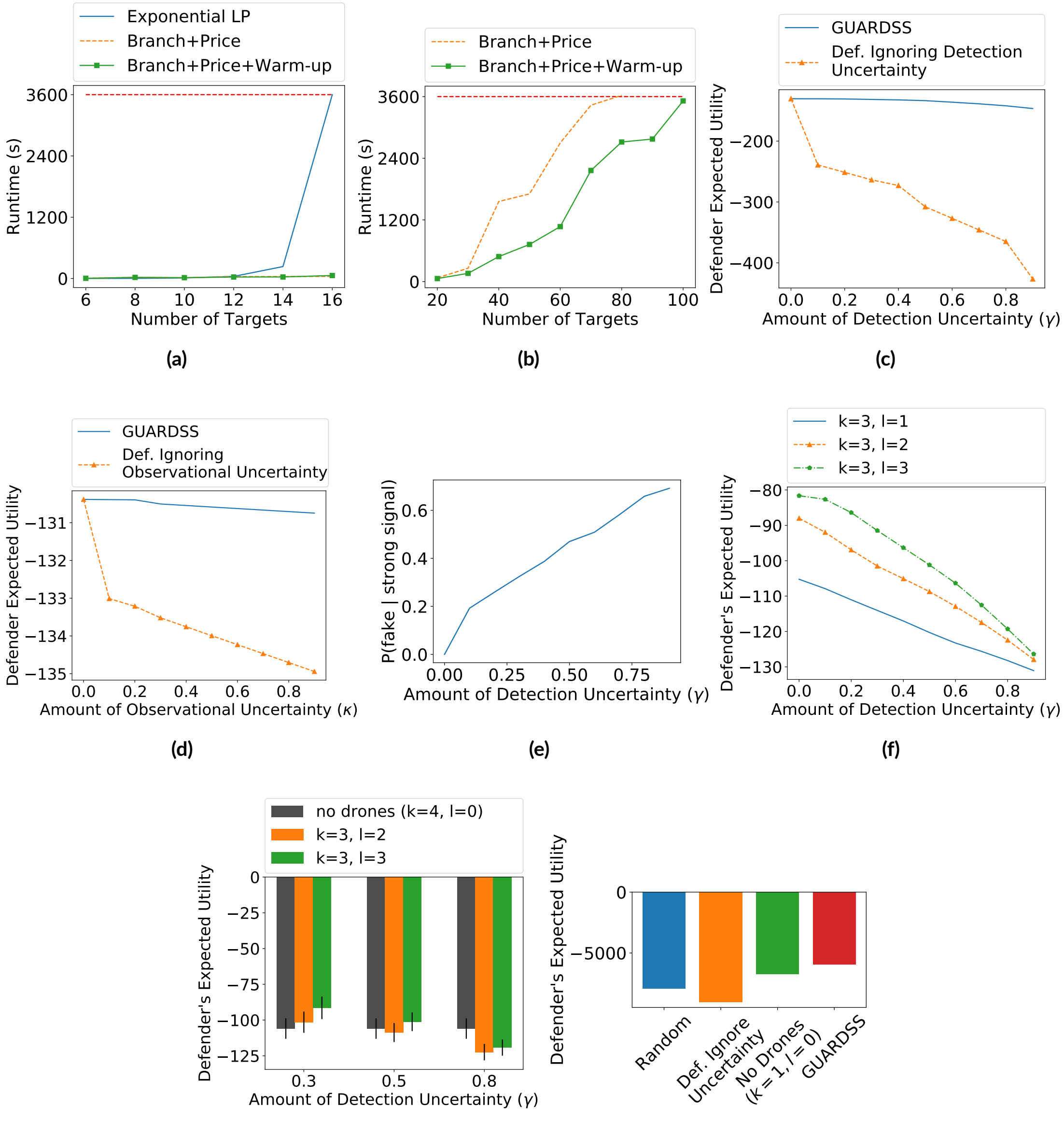
Despite considering uncertainty, sensors may be less valuable at a high level of uncertainty. In Fig. 1.3f, the defender expected utility is influenced by the number of drones and uncertainty in size $N = 15$ graphs. In Fig. 1.3g, drones are better than an extra patroller at $\gamma = 0.3$ ($p \leq 6.661e{-}02$), but at $\gamma = 0.8$, patrollers are better than drones ($p \leq 1.727e{-}07$).

1.7 Conservation Drones

We have deployed a drone in South Africa, equipped with a thermal camera and detection system[5]. A photo of the drone team in South Africa currently is included in Fig. 1.1 (center). To ease the challenges faced by these operators in coordination of drones with imperfect sensors and patrollers, we apply GUARDSS and show that it provides positive results in simulation to support future potential deployment. To facilitate the most realistic simulation possible, we utilize example poaching hotspots in a real park. We cannot provide the exact coordinates in order to protect wildlife, but we selected points based on geospatial features, and selected utilities to reflect the fact that the reward and penalty of the adversaries are impacted by animal presence, price, and distance to several park features used in[105]. The targets are shown in Fig. 1.4 (left). Any targets within 5 km are connected via edges in the graph, as park rangers could cover 5km for response. The resulting graph is shown in Fig. 1.4 (right). The utilities are included in the Appendix along with further details. For the simulation, we use 3 drones and 1 patroller. In the "no drones" scenario only, there are 0 drones and 1 patroller. We use $\gamma = 0.3$ for detection uncertainty and no observational uncertainty (see the Appendix for results with other γ). These details are directly input to GUARDSS, and then a mixed strategy is determined to cover the park. Fig. 1.3h shows the defender expected utility in this park using GUARDSS with and without uncertainty, and several baselines. A negative defender expected utility indicates that animals were lost, so a higher positive number is ideal. Therefore, we perform better with GUARDSS than using a random allocation, ignoring uncertainty, or forgoing drones. In fact, *ignoring uncertainty is worse than forgoing drones completely.* For varying γ (see Appendix), the gap between ignoring detection uncertainty and GUARDSS increases as γ increases, and the gap between the no drones case and GUARDSS decreases as γ increases, showing a similar trend to Fig. 1.3g. However, in all cases, the results emphasize the importance of correctly optimizing to get value from drones even with uncertainty.

1.8 Conclusion

The loss due to ignoring uncertainty in the initial data gathering stage can be high such that sensors are no longer useful. Nevertheless, by carefully accounting for uncertainty, uncertain information can still be exploited via a novel reaction stage and signaling even upon no detection. In this case, despite being aware of uncertainty, the adversary does not know whether he was detected, nor whether a patroller will respond in the reaction stage. Our results illustrate that the defender can exploit this informational advantage even with uncertain information. Thriving under this uncertainty makes real-world deployment of GUARDSS promising, as shown through simulation.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 1.3: Experimental results. Figs. 1.3a-1.3b compare multiple LPs approach (Exponential LP) with GUARDSS branch-and-price and heuristic method. Figs. 1.3c-1.3d show defender expected utility when amount of detection uncertainty γ and observational uncertainty vary. Defender expected utility decreases much more when uncertainties are ignored. Fig. 1.3e shows that the defender is leveraging the informational advantage as uncertainty increases. Figs. 1.3f-1.3g show that in the presence of a high false negative rate, extra patrollers may be more useful than drones. Fig. 1.3h contains the results from the case study, where GUARDSS performs best.

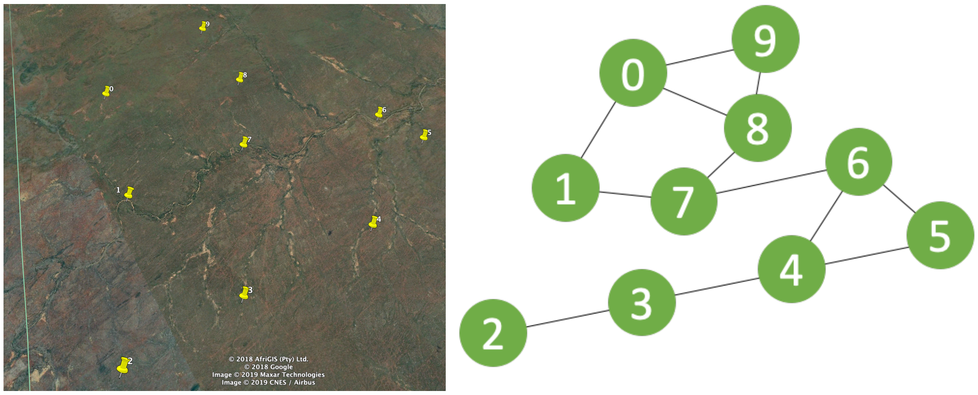

9
8
0
6
5
7
1
4
3
2
© 2018 AfriGIS (Pty) Ltd.
© 2018 Google
Image © 2019 Maxar Technologies
Image © 2019 CNES / Airbus
0
9
8
1
7
6
5
4
3
2

Figure 1.4: A park in Google Maps with potential poaching hotspots and the resulting graph (edges for $<$ 5 km).

2

Role of Human-AI Interaction in Selective Prediction

2.1 Introduction

Despite significant progress in machine learning-based AI systems, applications of AI to high stakes domains remains challenging. One of the challenges that arises is when humans interact with the

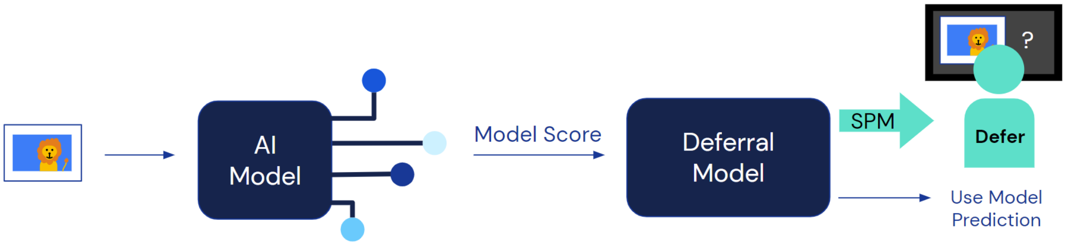


Figure 2.1: Deferral workflow: obtain AI model prediction for given input ("AI Model"); use model score as input to deferral model to decide whether to defer ("Deferral Model"); defer to humans if necessary, using SPM based on the prediction and deferral status ("Defer").

AI systems. For example, a human must assess the reliability of predictions made by a trained machine learning system, particularly when there is a distribution shift between the data the system was trained with and data encountered at deployment. In such situations, communicating the uncertainty associated with ML predictions appropriately is critical[159,117].

Given the difficulty of communicating probabilities to human users[99,340], a pragmatic alternative is to determine whether an AI system is more likely to make an erroneous prediction than a human, and defer to a human in such cases. A number of such settings have been studied in the literature, including selective prediction[313], learning to defer[208] and classification with a reject option[65]. While there are nuances that differentiate these works, in this paper, we will collectively refer to this body of literature as selective prediction and only emphasize the differences where relevant to our work. A related line of work considers human-AI teams in which humans receive AI assistance but make the final decision[19], and systems in which the AI makes the final decision with human input[314]. These prior works either: a) Assume that the human behaves identically even when they know that they are part of a human-AI team or b) Assume a utility-maximizing model for a human decision maker.

However, it has been documented that human-AI interaction may be more complex due to a range of cognitive phenomena. For example, humans have been shown to rely excessively on AI predictions (anchoring bias)[245,183], or even distrust AI predictions after observing AI mistakes[85]. Some work has begun to investigate solutions to these issues[54,245], however they have not focused on selective prediction systems, and context is critical.

In this work, we focus on binary classification tasks and study selective prediction systems (Fig. 2.1) that determine whether to rely on the outputs of an AI or defer to a human. To evaluate the overall performance of such a selective prediction system, it is important to model how the messaging, hereafter referred to as selective prediction messaging (SPM), that communicates the decision to defer impacts human accuracy. We run experiments with human subjects solving a challenging binary prediction task (that of detecting whether an animal is present in a camera trap image) and study the impact of different choices for communicating the deferring AI system's decision. We then perform statistical analysis on the human responses under various choices of SPM, and show that the choice of SPM significantly impacts human performance. Our results isolate two ingredients for a statistically effective communication strategy, that is, communicating that an AI system deferred (deferral status) and the AI system's predictions. Manipulating these leads to a boost in overall accuracy. We provide some plausible explanations for this phenomenon and suggest avenues for further work. Our contributions are therefore as follows:

- We develop and implement a balanced experimental design that can be used to measure the impact of SPM on the accuracy of selective prediction systems.
- We investigate the consequences of SPM on joint human-AI performance in a conservation prediction task, as opposed to prior work which assumes human behavior remains static during deferral.
- We discover two separable SPM ingredients, deferral status and prediction, which have distinct, significant effects on human performance, and demonstrate that manipulating these ingredients leads to improved human classifications in a human-AI team, implying that the

setup and the information given to humans during such tasks has a large impact on the performance of a human-AI team.

- We suggest that our results may relate to a more general property of naturalistic datasets, that in conditions that are ambiguous, sharing AI model predictions with humans can be detrimental.

2.2 Related Work

We start by describing prior work considering different potential roles in human-AI teams, namely, decision aids and systems in which an AI model defers to a human on challenging cases only. We then discuss supporting human-AI decision-making and selective prediction algorithms.

Decision Aids: Potential deployment scenarios for human-AI teams that have been discussed in the literature vary greatly depending on the application. One increasingly common scenario, particularly in high-risk domains, is that of AI systems serving as decision aids to humans making a final decision. For example, Green & Chen[115] explore the scenario of human decision-makers using a risk assessment model as a decision aid in financial lending and criminal justice (specifically, pretrial detention) settings. They found that humans were biased and that they failed to properly evaluate or take model performance into account, across different messaging conditions communicating the model's predicted risk. A decision aid for risk assessment is similarly studied in De-Arteaga et al.[80], particularly in the real-world domain of child maltreatment hotline screening. Humans indeed changed their behavior based on the risk assessment tool, but they were able to identify model mistakes in many cases. Gaube et al.[100] conduct experiments to measure the interaction between medical AI systems and clinicians. Radiologists who participated reported AI advice as lower quality than human advice, though all advice truly came from humans. Furthermore, clinician diagnostic accuracy was reduced when they were given incorrect predictions from the AI system. Even with humans making the final decision in each case, there is a great deal of variability between these AI systems, their impacts, and their application domains.

Deferral: AI systems have also played a slightly different role by making decisions in straightforward cases and deferring to human decision makers otherwise, which is most similar to our scenario. Wilder et al.[314] defer to an expert on cases that are best suited for human decision-making compared to model-based decision-making (determined using end-to-end learning), yet the final system is evaluated on historical human data, meaning humans did not know they were part of a human-AI team while labeling. Keswani et al.[154] propose deferral to multiple experts, including a classifier, by learning about the experts from their decisions. While Amazon Mechanical Turk was used to collect labels from participants to train and evaluate this model, it is similar to using historical human labels, as the humans again completed the task without knowing the deferral status or model prediction. Such deferral models are not yet widely deployed, nor have their impacts on humans been studied.

Supporting Human-AI Decision-Making: To mitigate some of the known negative impacts of AI on human decision-making, it may be beneficial to present humans with further information, such as uncertainty. Bhatt et al.[30] find that humans may unreliably interpret uncertainty estimates, but that the estimates may increase transparency and thereby performance. In a system used by people without a background in statistics, for example, presenting categories (such as deferral status) may make it easier to interpret. Bhatt et al.[30] advocate for testing this with humans, including of different skill levels and in different domains. Further suggestions for positive human-AI interaction are presented in Amershi et al.[10], including to "Show contextually relevant information," and "Scope services when in doubt." We are interested in finding the best way to implement these ideas with SPM.

Selective Prediction: Selective prediction can be traced back to the seminal work of Chow[65], where theoretical properties of classifiers that are allowed to "reject" (refrain from making a prediction) and ideal rejection strategies for simple classifiers are investigated. In these settings, the main metrics are the accuracy of the classifier on the non-rejected inputs and the rate of rejection, and the natural trade-off between these. A recent survey of theoretical work in this area can be found in Wiener & El-Yaniv[313]. We omit an extensive review of selective prediction literature, but acknowledge there is additional work in this area.

More relevant to our work is the work on learning to defer[192,208]. Madras et al.[192] propose to defer to a human decision maker selectively in order to improve accuracy and fairness of a base classifier. Mozannar & Sontag[208] develop a statistically consistent loss function to learn a model that both predicts and defers. Geifman & El-Yaniv[102] develop deferral strategies purely based on the confidence estimate of an underlying classifier.

None of these works model the impact of deferral (or its communication to a human decision maker) on the accuracy of a human decision maker. Since this is the primary object of study in our work, we do not include a full selective prediction literature review. We use a simple confidence-based deferral strategy inspired by the work of Geifman & El-Yaniv[102], but our experimental design is compatible with any selective prediction system.

2.3 Background

The primary goal of our work is to evaluate the impact of SPM on human performance in selective prediction systems. We design an experiment to evaluate this impact and base both our design and the questions we study on the psychological literature on joint decision-making.

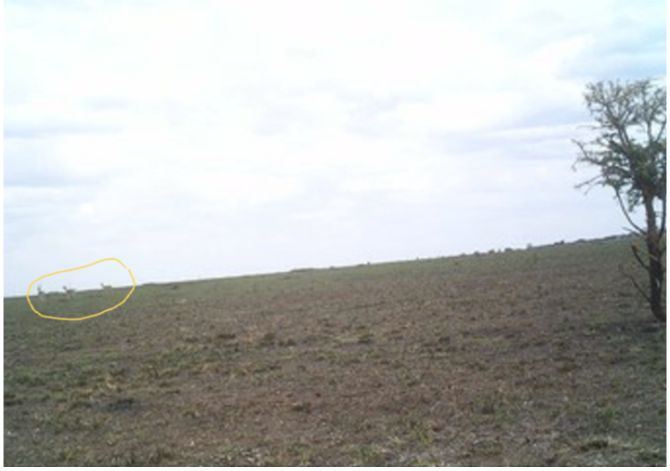


Figure 2.2: Example camera trap image with distant (circled) animals.

Psychology Literature on Human Decision-Making: The psychology literature has extensive studies on decision-making in human teams. Three psychological phenomena stand out as being relevant to our work: 1) Humans are sensitive to the specific way that a task is framed[298], 2) Humans are capable of flexibly deploying greater attentional resources in response to changing task demands and motivation[160], and 3) When deciding how best to integrate the decisions of others, humans take into account their own decision confidence as well as the inferred or stated confidence of other decision makers[45,18,195,14].

This suggests that SPM may impact how humans perceive their task, how much they trust the AI system, and ultimately how accurate their final prediction is, which is directly related to the composite performance of a selective prediction system. We consequently propose an experimental design where we take four natural choices for SPM and estimate their impact on the accuracy of human labelers.

Dataset: The dataset we use in this work is composed of images from camera traps, which are cameras triggered to capture images when there is nearby motion. These can be used to capture images of animals to understand animal population characteristics and even animal behavior, both of which are useful for conservation planning purposes. The volume of images generated in this manner is too high for manual inspection by rangers and scientists directly involved in conservation and monitoring efforts.

To alleviate this burden, the Snapshot Serengeti[*] project was set up to allow volunteers to apply rich labels to camera trap images[†]. These labels are publicly available[‡], and ground truth comes from label consensus from multiple individuals [282]. Given these labels, AI models have been developed that automatically classify and/or detect animals in camera trap images [222,284,26].

Whether relying on an AI model or volunteers, this processing is still difficult. Roughly seven out of ten images contain no animals, as they are the result of false triggers, e.g., due to heat and/or wind. However, it can be challenging to determine which images contain an animal at all, let alone the species, because of challenges like animal camouflage, distance to the camera, or partial visibility in the camera's field of view. An example camera trap image with animals on the horizon is shown in Fig. 2.2.

We consider a binary task where the bulk of blank images are first removed before images are uploaded to be labeled by volunteers or a species-identifying AI model. We investigate the use of a selective prediction model to filter out blank images, while prioritizing images for human review for which the blank/animal AI model is uncertain. This is similar to Willi et al. [315], Norouzzadeh et al. [221], as they also involve human-AI teams and remove blank images before species identification. However, 1) The human-AI teams differ, e.g., Norouzzadeh et al. [221] does not involve humans to remove blank images, but uses active learning for species identification, and 2) They do not focus on human behavior. To begin our workflow (Fig. 2.1), we obtain model scores from an ensemble of AI

[*] https://www.zooniverse.org/projects/zooniverse/snapshot-serengeti

[†] https://lila.science/datasets/snapshot-serengeti

[‡] https://lilablobssc.blob.core.windows.net/snapshotserengeti-v-2-0/zooniverse_metadata/raw_data_for_dryad.csv.zip

models that filters out blank images, then develop a deferral mechanism.

Deferral Mechanism: We use a selective prediction algorithm which finds the optimal threshold(s) for AI model scores to defer, as illustrated in Fig. 2.1. Concretely, consider a binary classification task with inputs x and labels $y \in \{0, 1\}$. We assume that we are given a pretrained ensemble of AI models m and have access to the predictions made by a human h as well. Given inputs x (for example, pixels of an image), we obtain a continuous score $m(x) \in [0, 1]$ that represents the confidence of the ensemble that the label corresponding to x is 1.

The deferral mechanism we use is a simple rule-based system that identifies ranges of the model score where the model is less likely to be accurate than a human and defers on these. In particular, we use a deferral model that identifies one continuous interval in the model score space to defer on.

The deferral model is parameterized by two real numbers $0 \leq \theta_1 \leq \theta_2$ and is defined as

$$\text{defer}(x; \theta) = \begin{cases} 1 \text{ if } m(x) \in [\theta_1, \theta_2] \\ 0 \text{ otherwise} \end{cases}$$

$\text{defer}(x) = 1$ represents the decision to defer on input x and $\text{defer}(x) = 0$ represents the decision to predict. Given θ, the AI model prediction x and the prediction made by a human $h(x)$, the selective prediction system is given by

$$\text{sp}(x; \theta) = \begin{cases} h(x) \text{ if defer}(x; \theta) \\ m(x) \text{ otherwise} \end{cases}$$

Given a dataset $\mathcal{D}$ of inputs, corresponding model scores, ground truth labels, and human labels, we choose θ by solving the following optimization problem:

$$\max_{\theta} \text{Accuracy}(\mathcal{D}; \theta) \text{ subject to DeferralRate}(\mathcal{D}; \theta) \leq r$$

where Accuracy$(\mathcal{D}; \theta)$ refers to the accuracy of the selective prediction system sp with parameters θ on the dataset, DeferralRate$(\mathcal{D}; \theta)$ refers to the fraction of points in the dataset for which defer$(x; \theta) = 1$, and r is a bound on the deferral rate, reflecting the acceptable level of human effort or budget constraints on hiring human decision makers.

This optimization problem can be solved in a brute force manner by considering a discrete grid on the $[0, 1]$ interval and going over all possible choices for the two thresholds θ.

Choosing a Deferral Model: We now describe how we choose a deferral model on the Serengeti dataset. Our goal is to find a model such that the accuracy of the sp classifier is higher than the human h or the AI ensemble m. A large fraction of the camera trap images are false positives, where the camera trap was triggered by a stimulus that was not an actual animal. In order to create a balanced dataset for tuning the deferral model, we subsample these empty images with no animal present. We implement a penalty for deferral for the sp classifier, which leads to varied performance at different deferral rartes, as seen in Fig. 2.3. In Fig. 2.3, individual human accuracy (as opposed to consensus accuracy, which is 1.0) is 0.961, and AI model accuracy is 0.972 (based on choosing one operating point to turn the continuous model scores into binary predictions that maximizes accuracy of the AI-only classifier). We also include the ideal performance if we had a perfect oracle to decide, for each image, whether to defer to a human or rely on the model (given historical human labels). This perfect selective prediction would achieve an accuracy of 0.994.

Given these results, we choose a deferral rate of 1% as an acceptable level of withholding (see Appendix for details) that still improves expected accuracy by a significant margin relative to AI-

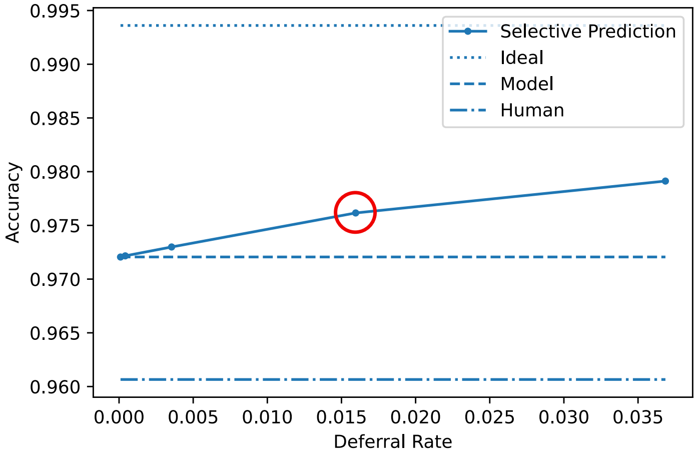
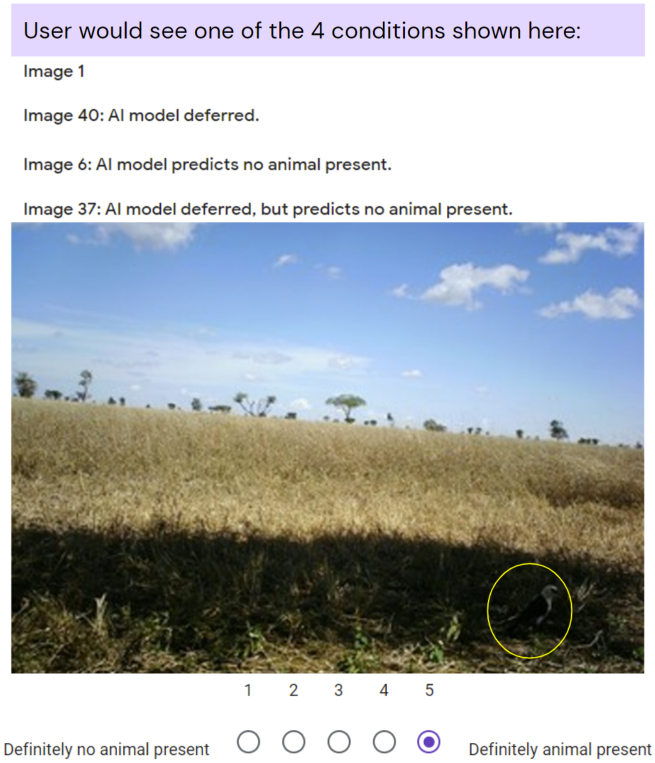


Figure 2.3: Tradeoff of expected accuracy and deferral rate. More deferral improves performance, but we still gain performance even when deferring less at the circled point.

only or human-only classifiers.

2.4 Experiment Design

We hypothesize that the accuracy of a human decision maker in a human-AI team is affected by the SPM in the last step of our workflow (Fig. 2.1). We specifically consider presenting information about the AI's **prediction** and **deferral status**. The AI prediction refers to the class returned by the AI model, e.g., animal or no animal. The deferral status refers to the result of selective prediction, in which we threshold the model score to determine whether to ask a human to review an image (defer), or rely on the AI prediction. We therefore design a human participant experiment with all possible combinations of these two details: 1) Neither message (NM), 2) Deferral status only (DO), 3) Prediction only (PO), and 4) Both messages (BM), as shown in Fig. 2.4. We create a survey to host this experiment, consisting of the following sections: 1) Information and consent, 2) Camera trap training and explanation, in which participants are introduced to camera trap images and given



User would see one of the 4 conditions shown here:

Image 1

Image 40: AI model deferred.

Image 6: AI model predicts no animal present.

Image 37: AI model deferred, but predicts no animal present.

	1	2	3	4	5	
Definitely no animal present	○	○	○	○	●	Definitely animal present

Figure 2.4: The four possible SPM conditions in our experiment, along with a challenging example image. The animal is circled in the image here for the reader's convenience, but in our experiment the circle was not present.

an example with an explanation for the best label, 3) Adding AI assistance, in which participants are told about adding an AI model and deferral to assist in the task of sorting camera trap images, 4) AI assist practice and explanations, which consists of 10 examples (9 correct, 1 incorrect) drawn from the Serengeti validation set with AI model assistance, 5) Post training questions asking participants to describe the model strengths and weaknesses, 6) Labeling, and 7) Post dataset questions. Several of these design choices align with the guidelines from Amershi et al. [10], including describing the AI performance, and providing examples and explanations.

In the labeling section, we display 80 model-deferred images (like Fig. 2.4) under the four SPM conditions (yielding 20 images per condition, which we believe balances a reasonable number of examples with a manageable amount of participants' time). Each includes a request for labels of animals present or not, with a Likert scale as in Fig. 2.4. The images are randomly allocated across the four communication conditions. **We did not inform participants that all images were deferred, we only relied on the different SPM conditions.** Each participant judged images across all four conditions, and no single image was presented more than once to the same participant. The set of 80 images are sampled so as to balance the number of true positive, false positive, true negative, and false negative model classifications, and therefore additionally ensure a balance in the number of images across classes. To ensure that there are no effects due to the specific order or allocation to a condition of each image, four separately seeded random allocations are carried out and each participant is randomly assigned to one of them. To test if the effects of the experimental conditions on accuracy exceed variation expected by chance, the data are analysed in a 2x2x2 within-subject repeated measures ANOVA with the factors "deferral status" (with the levels: "shown" and "not shown"), "prediction" (with the levels: "shown" and "not shown") and "model accuracy" (with the levels: "model correct" and "model incorrect"). We received approval from an internal ethics review board, and then recruited 198 participants from Prolific to take part in the experiment. Responses from all 198 participants were included in the ANOVA. Aggregated data are available at https://github.com/deepmind/HAI_selective_prediction/.

2.5 Experiment Results and Analysis

2.5.1 SPM Affects Human Accuracy

As can be seen in Fig. 2.5, the information provided to human participants about the model affected participants' accuracy. The human-AI communication method that yields the highest human performance is DO. Accuracy in this condition is significantly greater than either humans classifying images by themselves without any information about the model (mean of DO: 61.9%, mean of NM: 58.4%, $p < 0.001$), or the model operating alone (mean of DO: 61.9%, mean of model alone:

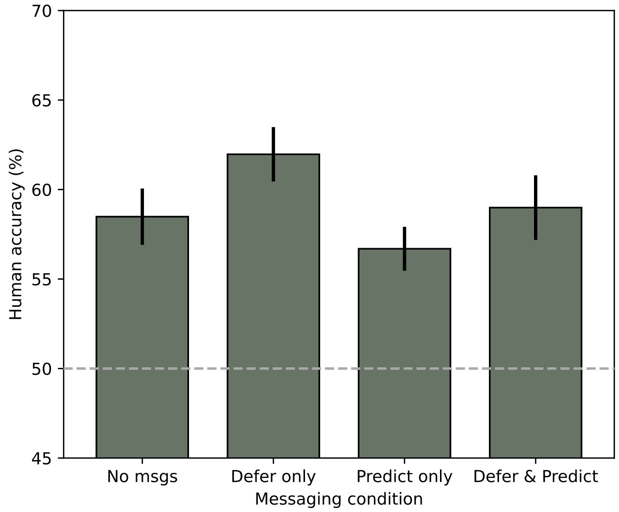


Figure 2.5: Accuracy of human participants on deferred images, across different SPM conditions. Each bar shows the participant classification accuracies across the entire dataset, errorbars show 95% confidence intervals on the mean. Participants' responses are more accurate when the images are accompanied by the context that the images are deferred (DO and BM vs. NM and PO). Showing the model's prediction of the label has a negative effect on accuracy. The horizontal dashed grey line indicates chance performance (50%).

50%, $p < 0.001$). Model performance is 50% since the images presented to humans are subsampled from the set of model-deferred images. Furthermore, across all responses, participants are significantly more accurate when the deferral status is shown (conditions DO and BM) than when it is not (conditions PO and NM) (mean of deferral status: 60.4%, mean of no deferral status: 57.4%, $p < 0.001$). We believe this effect of deferral status may be driven by participants inferring that the images are likely to be quite difficult, and therefore concentrating harder. By contrast, participants are significantly less accurate when the prediction is shown (mean of prediction: 57.8%, mean of no prediction: 60.2%, $p = 0.003$). Overall, therefore, we find that providing the deferral status, while avoiding the provision of model predictions, leads to the highest accuracy in human decision-making in this context.

2.5.2 Model Predictions Influence Human Decisions

While the preceding results clearly suggest that participants use the model predictions at least some of the time, the evidence is indirect. We therefore conduct a more direct analysis targeting this question. Specifically, for each image, we compute the proportion of raters who agree with the model prediction under 1) NM and 2) PO. As each image is presented under both conditions, we subtract these two scores to assess how much humans increase their agreement with the model when the model prediction is present. We refer to this as the "conformity" score, and plot it in Fig. 2.6. Across the set of images, we find an average conformity score of 0.08 (i.e., raters are influenced on 8% of trials), which is significantly greater than zero ($p < 0.0001$), demonstrating that raters use the model information when present. We further test a hypothesis suggested by Bahrami et al.[18], Boorman et al.[45], that people are more likely to use model information when they are less confident in their own decisions (as measured by Likert ratings in this case). We separately compute the conformity score for low and high confidence human decisions, and find that, as expected, conformity is higher when rater confidence is low (mean of low confidence: 0.116, mean of high confidence: 0.045, $p = 0.014$).

2.5.3 Model Accuracy Affects Human Accuracy

While the preceding analysis demonstrates that people do indeed use the model predictions, this in itself does not explain why we find a decrease in human accuracy when model predictions are available. This suggests a potential bias, such that people tend to use the model information more when it is actually incorrect. To explore this possibility, we focus on the effect of the provided SPM only on the images that the model classifies incorrectly (Fig. 2.7). We observe that participants' accuracy in the PO condition is significantly reduced. While participants perform at chance in the other conditions, participants perform significantly below chance in the PO condition (mean of other

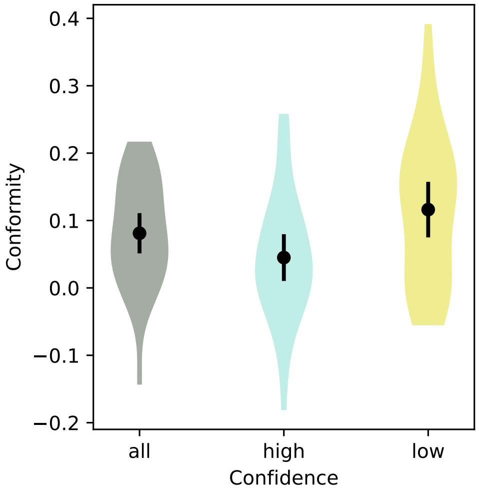


Figure 2.6: Conformity is the relative increase in agreement when the model predictions are present. For each image, we show how much the agreement between the set of human predictions and model prediction changes from the NM to the PO conditions. Conformity is significantly higher when humans are less confident, indicating that humans are influenced by model predictions more when they are less certain about their judgement.

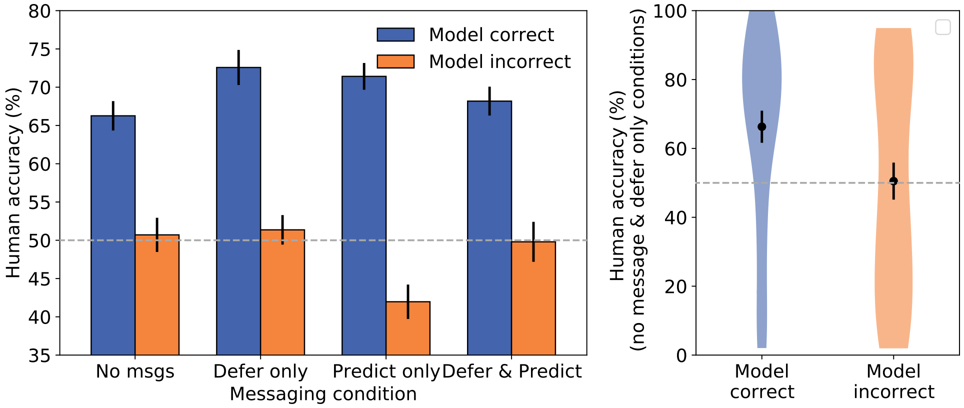


Figure 2.7: Accuracy of human participants on deferred images, split by whether the model correctly or incorrectly classified the associated images. (Left) Results split by whether the model correctly classified the image. Notably, images where the model is incorrect has lower participants' accuracy across all conditions, generally at chance (dashed grey line). Crucially, participants are significantly below chance in the PO condition, for which the model prediction misleads humans. (Right) Human accuracy for each image in the NM and DO conditions, split by whether the model labels the image correctly. Participants' accuracy is significantly reduced (to around chance) for the subset of images which the model fails to label correctly. Together, both show a congruence between what images the model and humans find difficult to label correctly.

conditions: 50.6%, mean of PO: 41.9%, $p < 0.001$). In this condition, the label provided by the model is most salient, and critically it provides wrong information. Participants appear to integrate this information as they perform 8.7% worse in the PO condition compared to the NM condition, on the images that are misclassified by the model. In contrast, on images that are correctly classified by the model, and therefore the model can provide a correct prediction message, participants gain 5.1% of accuracy between the NM and PO conditions. This asymmetry, that incorrect model prediction messages are integrated more by the participants than correct model prediction messages, appears to lie at the heart of why we observe an overall negative effect of the prediction message. This pattern is consistent with the observation that participants and the model tend to err on the same images, as shown in the violin plots of Fig. 2.7. Specifically, participants are correct 66.3% of the time on images shown in the NM and DO conditions and that the model correctly classifies, but only 50.5% on images in these same conditions that the model classifies incorrectly ($p < 0.05$). Additionally, there is a positive correlation between average human Likert scores in the NM and DO conditions and model scores on the same images, suggesting that humans and models learn similar semantic task dimensions (Pearson's $r = 0.27, p = 0.021$).

2.5.4 Asymmetry in Human-AI Agreement

Our results demonstrate a differential change in human accuracy in the presence of model information, based on whether the model is correct or incorrect. When we subdivide the data based on whether the model is correct or incorrect, we see that human accuracy tends to be lower when the model is incorrect (Fig. 2.7, left). Additionally, exposing humans to incorrect predictions of an AI makes them even more likely to be incorrect. We confirm that human accuracy tends to be lower when the model is incorrect by directly computing the image-wise agreement between human and model ratings on images where the model is correct vs. incorrect (Fig. 2.7, right). This analysis is performed only on the trials under the NM condition, as we want to investigate independent agreement. As expected, we find that agreement is significantly higher for the correctly labelled images (mean agreement of model correct: 69.6%, mean of model incorrect: 44.9%, $p = 0.007$). This difference in prior agreement over images has potentially important repercussions when introducing model predictions. Specifically, as the subset of images where the model is correct already has high prior agreement with the human raters, there is much less potential for the model to influence human judgments, and hence increase accuracy. By contrast, because the images that are misclassified by the model have a lower prior agreement level with human raters, there is greater potential for model influence, which in this case will decrease accuracy. This asymmetry in potential model influence between correct and incorrect trials is likely to account for the overall drop in human accuracy we see when model predictions are provided to human raters. This is an important result, as it demonstrates that the specific pattern of covariance between model and human decisions can lead to significant downstream effects on joint decision-making when humans have access to the model predictions.

2.6 Discussion

To summarize, we find that there are significant effects in the way deferred images are presented to humans in a selective prediction workflow. In particular, in this context, presenting deferral status is helpful, while presenting the uncertain prediction, even when accompanied by a deferral status, can be harmful, especially when the model is incorrect. From this, it is clear that performance will not necessarily be what is estimated from historical human labels. Together, these findings illustrate the importance of considering the human-AI team while designing selective prediction systems, as the SPM can have a significant impact on performance.

We believe these are important findings to direct future research, and have several suggestions for open questions to explore. First, this experiment could be expanded. While we chose to leave this static, it would be helpful to determine the amount and type of training that is most useful for participants. We also chose to use two categories for deferral status, either defer or not defer. However, it is possible that finer-grained information about model uncertainty could be helpful[30]. We additionally focused on understanding why prediction hurt, but encourage collecting further information, such as timing, to better understand why deferral status helped. This may help inform further research into designing selective prediction algorithms based on human-AI teams, for example by exploring bounded rationality for training improved selective prediction models.

There are also questions about generalizability of these results. These specific results (i.e., that deferral status helps while prediction hurts performance) are not likely to be robust across datasets, different human-AI use scenarios (e.g., decision aids), or even participant expertise levels. For example, in this study, we asked two domain experts working with the Serengeti dataset to go through the same survey provided to Prolific participants. We similarly find that each individual has different performance in the four conditions, and that the deferral status leads to improved performance in both cases. However, the interactions are slightly different. It is necessary to search for generalizable trends across these cases in future work.

Finally, though we worked with humans in this study, it is extremely important to consider specific deployment challenges in these contexts, such as how selective prediction may change existing processes, e.g., in healthcare[332], or how to best modify the workflow and instructions in the case where there are multiple human experts we could rely on. In all cases, we highlight the promise of human-AI teams, but stress the importance of remembering the human component of human-AI teams.

3

VIOLA: Video Labeling Application for Security Domains

3.1 Introduction

Game-theoretic approaches have led to applications that have been successfully deployed in infrastructure security domains such as protecting airports, ports and metro systems[287], as well as in green security domains such as protecting wildlife, forests, and fisheries[93,140,122]. In these game-theoretic approaches and security applications, input data are needed to determine the payoff structure of the game, to learn the behavioral models of the players, and to predict where adversaries are more likely to attack. In previous efforts, the data were provided by domain experts directly[238], recorded by practitioners in the field over months or years[217,149], or collected through human subject experiments on platforms such as Amazon Mechanical Turk (AMT)[148].

Videos taken by drones have become an emerging source of massive data[127], especially in the domain of wildlife protection (e.g., the PAWS security games application[93]). For example, detecting wildlife from conservation drone videos can help estimate the animal distribution density, which decides the payoff structure of the game. Detecting humans and their movement patterns could lead to successful learning of adversaries' behavioral models, which is an important topic in security games[219,148]. Data collected from conservation drones can not only be used to provide input data to the game-theoretic models, but can also enable the development of a new generation of game-theoretic tools for security. The data can be used to train or fine-tune a deep neural network to automatically detect adversaries from the video taken by the conservation drones in real-time.

Unfortunately, collecting labeled data from videos taken by conservation drones can be a labor-intensive, time-consuming task. To our knowledge, there is no existing application that focuses on assisting in the labeling of videos taken by conservation drones in security domains. Existing applications for labeling images[84,90] cannot be directly applied to labeling videos, as treating each frame as a separate image can lead to inefficiency since it does not exploit the correlation between frames. Video labeling applications such as VATIC[305] attempt to choose key frames for labeling, or track objects through the video. However, in conservation drone videos with camera motion, possibly collected using a different wavelength, these methods may not apply and may lead to inaccurate results or extra work for labelers, since the position of the objects in the video may change abruptly and the lack of color bands makes the tracking much more difficult. Furthermore, these applications are often paired with AMT to get labeled video datasets from online workers. However, in a security domain with sensitive data, meaning data that would provide adversaries with some knowledge of defenders' strategies should it be shared, it may be undesirable to use AMT. This would then require finding labelers, and setting up an internal system to keep the process organized.

In this paper, we focus on better collection of labeled data from conservation drones to provide input for game-theoretic approaches for security, and in particular to security game applications for wildlife conservation such as PAWS[93]. There has been work on labeling tools in domains such as computer vision and cyber security[84,57], but there exists no work on labeling tools for game-theoretic approaches in security domains. Most previous work on game theory for security ignores where the payoffs and behavioral models come from, and we fill the gap.

In particular, we will focus on labeling videos taken by long wave thermal infrared (hereafter referred to as thermal infrared) cameras installed on conservation drones, in the domain of wildlife security. We present VIOLA (VIdeO Labeling Application), a novel application that assists labeling objects of interest such as wildlife and humans. VIOLA includes a workload distribution framework to efficiently gather human labels from videos in a secured manner. We distribute the work of labeling the videos and reviewing the labels amongst a small group of labelers to ensure efficiency and data security. VIOLA also provides an easy-to-use interface, with a set of features designed for conservation drone videos in the wildlife security domain, such as allowing for moving multiple bounding boxes simultaneously and tracking bright spots in the video automatically. We will also discuss the various stages of development to create VIOLA, and we will analyze the impact of different labeling procedures and versions of the labeling application on efficiency, with a particular emphasis on the surprising results that showed some changes did not increase the efficiency.

3.2 Related Work

Game-theoretic approaches have been widely used in infrastructure and green security domains[287]. In green security domains such as protecting wildlife from poaching, multiple research efforts in artificial intelligence and conservation biology have attempted to estimate wildlife distribution and poacher activities[93]; such efforts often rely on months or years of recorded data[217,149]. With the recent advances in conservation drone technology, there is an opportunity to provide detailed data about wildlife and poachers for game-theoretic approaches. Since a poacher is rewarded for successfully poaching wildlife, the wildlife distribution determines the payoff structure of the game. Poachers' behavioral models can be inferred from poaching activities and be used to design better patrol strategies with game-theoretic reasoning. In addition, game-theoretic patrolling with alarm systems[9,23] has been studied. Conservation drones can provide input for such systems in real-time using computer vision, particularly by detecting humans in the conservation drone videos.

Detecting adversaries in the conservation drone videos is related to object detection. Recently, great progress has been achieved in computer vision by deep learning in object detection and recognition[249,247]. However, state-of-the-art detectors cannot be directly applied to our aerial videos because most methods focus on detection in high resolution, visible spectrum images. An alternative approach to this detection is to track moving objects throughout videos. Tracking of both single and multiple objects in videos has been studied extensively[335]. These methods also rely on high resolution visible spectrum videos. Single object trackers use discriminant features from high resolution videos to establish correspondences[162]. Much of multi-object tracking research is directed towards pedestrians[17,341,204], and primarily focuses on visible spectrum videos with high resolution, or videos taken from a fixed camera (except[204]).

Simpler and more general tracking algorithms exist that do not necessarily have these dependencies, such as the Lucas-Kanade tracker for optical flow[186], popular in the OpenCV package, and general correlation-based tracking[189]. Small moving objects can also be detected by a background subtraction method after applying video stabilization[230]. Because these methods are more general, they are still applicable to our domain and were explicitly tested, but still did not perform well in many cases. For example, since the video stabilization and background subtraction method assumes a planar surface, in the case of more complex terrain, there were many noisy detections. Instead of using tracking for detection, we therefore decided to focus on deep learning.

In order to use deep learning-based detection methods with aerial, thermal infrared data, hand-labeled training data are required to fine-tune the networks or even train them from scratch. In addition to video labeling applications such as VATIC[305], there has been work on semi-automatic labeling[330] and label propagation[16] which combines the effort of human labelers and algorithms to speed up the labeling process for videos. This work often focuses on how to select the frames for human labelers to label and how to propagate the labels for the remaining frames. This is difficult for our domain because of the motion of conservation drones, and because it is often hard for humans to tell which objects are of interest without seeing the object's motion. As a result, we sought to develop our own labeling application, VIOLA. The first key component of the application is a workload distribution framework. A common framework for image and video labeling is a majority voting framework[216,233,220,268]. VIOLA uses a framework based upon[90] to efficiently gather labels from a small group of labelers. We examine the framework further in Sec. 3.6 and Sec. 3.7.

3.3 Domain

Conservation drones are able to cover more ground than a stationary camera and can provide the defenders more advanced notice of a potential threat. To detect human activities at night, the conservation drones can be equipped with thermal infrared cameras. This is the type of conservation drone video we deal with in our domain, since poaching often occurs at night. We will specifically be able to use these types of data to detect humans and provide advanced notice to park rangers, and use these detections to provide input for patrol generation tools such as PAWS.

In order to accomplish this, we need labeled data from the thermal infrared, conservation drone videos in the form of rectangular "bounding boxes" for objects of interest (animals and humans) in each frame, with a color corresponding to their classification. However, the movement of conservation drones and the thermal infrared images make it extremely difficult to label videos in this domain. First, thermal infrared cameras are low-resolution, and typically show warmer objects as brighter pixels in the image, although the polarity could be reversed occasionally. Different phenomena could also cause brighter pixels without a warm object. For example, the ground warms during the day, and then emits heat at night, which can be reflected under a tree canopy and lead to an amplified signal that might look like a human or animal. Furthermore, vegetation often looks bright and similar to objects of interest, as in Fig. 3.1, where there are three humans labeled with bounding boxes, amongst many other bright objects. Second, since the data are captured aboard a moving conservation drone, these data often vary drastically. For example, the resolution, and therefore size of targets, is very different throughout our dataset because the conservation drone flies at varying altitudes.

In addition to difficult, variable video data to begin with, some videos may have many objects of interest in them, whereas some videos may not have any objects of interest at all. It sometimes takes a long time to determine if there are any objects of interest, and it also often takes a long time to label when there are many objects of interest. To illustrate the variation in the number of objects of interest, we analyze the historical videos we get from our collaborator. Fig. 3.2 shows a histogram of the average number of labels per frame, meaning that all frames in the video were counted, regardless of whether or not they were labeled, and a histogram of the average number of labels per labeled frame, meaning only frames that had at least one label were counted.

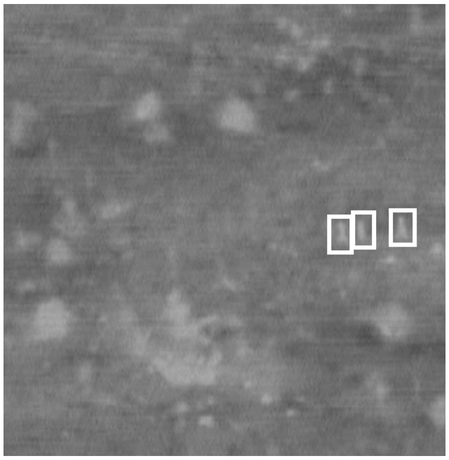


Figure 3.1: An example of a thermal infrared frame, where the three humans outlined by the white boxes look very similar to the surrounding vegetation.

3.4 Example Game-Theoretic Uses

We now provide two more specific examples of game-theoretic approaches that may be derived from the data acquired using VIOLA. First, we focus on using the labeled data directly for behavioral models. Second, we discuss using the labeled data to train deep learning models for further data analysis.

With the labels provided by VIOLA and information about each frame, such as GPS and camera angle, we can locate humans exactly throughout labeled videos. As such, we likely know the location of poaching activities and could use this information to learn how poachers make decisions on where to poach. In particular, we could use an existing behavioral model, such as SUQR[219], and the location of poaching activity derived from the labels to update or improve the behavioral model for poachers, which would better inform patrol strategies. Furthermore, we could analyze the movement of poachers, and a new behavioral model could be built using these movement patterns, in which poachers could choose a path instead of simply choosing a target to attack. This new behavioral model could be exploited to plan game-theoretic patrols.

In addition to directly using the labels from VIOLA for behavioral models, the labels could be used to train a deep learning model to automatically identify humans in real-time video streams. Similarly, we could use the output from the deep learning algorithm for behavioral models, and the automated identification would allow us to circumvent the need for human labelers when incorporating data collected in the future into the behavioral models. Moreover, patrollers could make online decisions during patrols without the need for additional personnel to monitor the videos in the field. The ability to make online decisions during patrols could lead to new models of game-theoretic patrolling. Patrols could even be made for the conservation drones themselves, which could introduce some behavioral challenges. The conservation drones could also potentially be used as a deterrent, so flying conservation drones could serve to both detect and deter poaching activities, while also collecting more data. In short, VIOLA has the potential to provide data that will better inform behavioral models and patrollers in the field, and introduce new questions that can be answered using game-theoretic approaches.

3.5 VIOLA

The main contribution of this paper is VIOLA, an application we developed for labeling conservation drone videos in wildlife security domains. VIOLA includes an easy-to-use interface for labelers and a basic framework to enable efficient usage of the application. In this section, we first discuss the user interface and then the framework for work distribution and training process for labelers.

3.5.1 User Interface of VIOLA

The user interface of VIOLA was written in Java and Javascript, and hosted on a server through a cloud computing service so it could be accessed using a URL from anywhere with an internet connection.

Before labeling, labelers were asked to login to ensure data security (Fig. 3.3a). The first menu that appears after login (Fig. 3.3b) asks the labeler which mode they would like, whether they would like to label a new video or review a previous submission. Then, after choosing "Label", the second menu (Fig. 3.3c) asks them to choose a video to label. Fig. 3.4 is an example of the next screen used for labeling, also with sample bounding boxes that might be drawn at this stage. Along the top of the screen is an indication of the mode and the current video name, and along the bottom of the screen is a toolbar. First, in the bottom left corner, is a percentage indicating progress through the video. Then, there are four buttons used to navigate through the video. The two arrows move backwards or forwards, the play button advances frames at a rate of one frame per second, and the square stop button returns to the first frame of the video. The next button is the undo button, which removes the bounding boxes just drawn in the current frame, just in case they are too tiny to easily delete. Also to help with the nuisance of creating tiny boxes by accident while drawing a new bounding box or while moving existing bounding boxes, there is a filter on bounding box size. The trash can button deletes the labeler's progress and takes them back to the first menu after login (Fig. 3.3b). Otherwise, work is automatically saved after each change and re-loaded each time the browser is closed and re-opened. The application asks for confirmation before deleting the labeler's progress and undoing bounding boxes to prevent accidental loss of work. The check-mark button is used to submit the labeler's work, and is only pressed when the whole video is finished. Again, there is a confirmation screen to avoid accidentally submitting half of a video. The copy button and the slider will be described further in Sec. 3.6. The eye button allows the labeler to toggle the display of the bounding boxes on the frame, which is often helpful during review to check that the labels are correct. Finally, the question mark button provides a help menu with a similar summary of the controls of the application (Fig. 3.5). Notice the bounding boxes surrounding the animals in this video are colored red. Humans would be colored blue. This is also included in the help menu.

To draw bounding boxes, the labeler can simply click and drag a box around the object of interest, then click the box until the color reflects the class. Deleting a bounding box is done by pressing SHIFT and click, and selecting multiple bounding boxes is done by pressing CTRL and click, which allows the labeler to move multiple bounding boxes at once. Finally, while advancing frames, bounding boxes drawn in the current frame are moved to the next frame. It only happens the first time a frame is viewed since it could otherwise add redundant bounding boxes or replace the bounding boxes originally added by the labeler.

If "Review" is chosen in the first menu after login, the second menu also asks the labeler to choose a video to review, and then a third menu (Fig. 3.3d) asks them to choose a labeling submission to review. It finally displays the video with the labels from that particular submission, and they may begin reviewing the submission. The two differences between the labeling and review modes in the application are (i) that the review mode displays an existing set of labels and (ii) that labels are not moved to the next frame in review mode.

3.5.2 Use of VIOLA

Our goal in labeling the challenging videos in the wildlife security domain is first to keep the data secure, and second, to collect more usable labels to provide input for game-theoretic tools for security. In addition, we aim for getting exhaustive labels with high accuracy and consistency. To achieve these goals, we distribute the work among a small group of labelers in a secured manner, assign labelers to either provide or review others' labels, and supply guidelines for the labelers.

Distribution of Work To keep the data (historical videos from our collaborators) secure, instead of using AMT, we recruit a small group of labelers, in this work 13. Labelers are given a username and password to access the labeling interface, and the images on the labeling interface cannot be downloaded.

In order to achieve label accuracy, we use a framework of label and review. The idea is simply that one person labels a video, and another person checks, or reviews, the labels of the first person. By checking the work of the labeler, the reviewer must agree or disagree with the original set of labels instead of creating their own. Upon disagreement, the reviewer can change the original labels. This was primarily chosen because it was clean, leading to one set of final labels.

We use spreadsheets to share both assignments and completion progress with the team of labelers. We ask labelers to include the time it took for them to complete their assignment in order to help make future assignments more reasonable in terms of time commitment, and in order to track the efficiency and success of the application itself. In addition, we split long videos into segments to make it easier to respect labelers' time commitments, and to finish extremely long videos quickly. There are also some videos that have long periods of nothingness, which are easier to ignore when the video is split.

Guidelines and Training for Labelers In order to achieve accuracy and consistency of labels, we provide guidelines and training for the labelers. During the training, we show the labelers several examples of the videos and point out the features of interest. We provide them with general guidelines on how to start labeling a video, as below.

In general, the process for labeling should be:

- Watch the video once all the way through and try to decide what you see.
- Once you have an idea of what is happening in the video by going through it, return to the beginning of the video and start labeling.
- Make and move bounding boxes.
- Send screenshots (including the percentage in the videos) if you need help.

In general, the process for reviewing should be:

- Refer to the guidelines and special circumstances directions.

- Go through the video, and use the eye button to check the original labels.
- Move, create, or delete bounding boxes as necessary, either as you go or after watching the whole video. Try not to resize the bounding boxes unless they are much too big or too small. Only change the classification and add or delete boxes if certain, and please confirm with us if not.
- Send screenshots (including the percentage in the videos) if you need help.

We also provide special instructions for the videos in our domain of interest, including a few key clues. For example, animals tend to be in herds, obviously shaped like animals, and/or significantly brighter than the rest of the scene, and humans tend to be moving. We also provide the following additional guidelines.

Directions for special circumstances:

- Only label when objects are bright since the polarity changes occasionally
- If something is occluded completely: do not label
- If something is occluded but you can still see most features of them: label
- If something is shaped like a human but never moves: do not label
- If something is cutoff halfway in/out of the frame: do not label
- If there are "ghosts" (Fig. 3.6): do not label
- If you cannot recognize an individual (i.e., distinct humans and animals): do not label

The final instruction about distinct objects is one of the more difficult instructions to follow in practice because often, the aerial view and small targets make it difficult to tell if there are one or more animals. The movement instruction is also difficult, since with so few pixels on objects plus camera motion, it sometimes looks like objects are moving that are not. In these ambiguous cases, labelers are encouraged to seek help. In cases of disagreement after discussion, we err on the side of caution and only label certain objects.

3.6 Development

Thanks in large part to feedback provided by the labelers, we were able to make improvements throughout the development of the application to the current version discussed in Sec. 3.5.1. In the initial version of the application, we had five people label a single video, and then automatically checked for a majority consensus among these five sets of labels. We used the Intersection over Union (IoU) metric to check for overlap with a threshold of 0.5 [90]. If at least three out of five sets of labels overlapped, it was deemed to be consensus, and we took the bounding box coordinates of the first labeler. Our main motivation for having five opinions per video was to compensate for the difficulty of labeling thermal infrared data, though we also took into account the work of[216] and [233]. The interface of the initial version allowed the user to draw and manipulate bounding boxes, navigate through the video, save work automatically, and submit the completed video. Boxes were copied to the next frame and could be moved individually. To get where we are today, the changes were as listed in Table 3.1.

The most significant change made during the development process was the transition from five labelers labeling the same video and using majority voting to get the final labels (referred to as “MajVote”) to having one labeler label the video followed by a reviewer reviewing the labels (referred to as “LabelReview”). We realized that having five people label a single video was very time consuming, and the quality of the labels was still not perfect because of the ambiguity of labeling thermal infrared data, which led to little consensus. Furthermore, when there was consensus, there were three

Table 3.1: Changes made throughout development.

Version	Change	Date of Change	Brief Description
1	-	-	Draws and edits boxes, navigates video, copies boxes to next frame
2	Multiple Box Selection	3/23/17	Moves multiple boxes at once, to increase labeling speed
3	Five Majority to Review	3/24/17	Requires only two people per video instead of five to improve overall efficiency
4	Labeling Days	4/12/17	Has labelers assemble to discuss difficult videos
5	Tracking	6/17/17	Copies and automatically moves boxes to next frame

to five different sets of coordinates to consider. Switching to LabelReview eliminated this problem, providing a cleaner and also time-saving solution. Another change, "Labeling Days", consisted of meeting together in one place for several hours per week so labelers were able to discuss ambiguities with us or their peers during labeling. Finally, the tracking algorithm (Alg. 1) was added to automatically track the bounding boxes when the labeler moves to the new frame. The goal was to improve labeling efficiency, as the labelers would be able to label a single frame, then simply check that the labels were correct.

An example of the tracking process in use is shown in Fig. 3.7. First, the labeler drew two bounding boxes around the animals (Fig. 3.7a), then adjusted the search size for the tracking algorithm using the slider in the toolbar (Fig. 3.7b). The tracking algorithm was applied to produce the new bounding box location (Fig. 3.7c). In contrast, the copy feature, activated when the copy button was selected on the toolbar, only copied the boxes to the same location (Fig. 3.7d). In this case, since there was movement, and the animals were large and far from one another, the tracking algorithm correctly identified the animals in consecutive frames. If several bright objects were in the search

Algorithm 1 Basic Tracking Algorithm

```
$bufferPixels \leftarrow userInput$
for all $boxesPreviousFrame$ do
    if $boxSize > sizeThreshold$ then
        $newBoxCoordinates \leftarrow boxCoordinates$
    else
        $searchArea \leftarrow newFrame[boxCoordinates + bufferPixels]$
        $thresholdedImage \leftarrow$ THRESHOLD(searchArea, threshold)
        $components \leftarrow$ CONNECTEDCOMPONENTS(thresholdedImage)
        if $numberComponents > 0$ then
            $newBoxCoordinates \leftarrow$ GETLARGESTCOMPONENT($components$)
        else
            $newBoxCoordinates \leftarrow boxCoordinates$
        end if
    end if
    COPYANDMOVEBOX($newFrame$, $newBoxCoordinates$)
end for
```

region, it could track incorrectly and copying could be better. One direction of future work is to improve the tracking algorithm by setting thresholds automatically and accounting for close objects.

3.7 Analysis

In this section, we analyze how the changes we made during the development of VIOLA affect labeling efficiency. To do this, we examine two questions: (i) how the changes affect the overall efficiency of the data collection process, which is measured by the total person time needed to get a final label – a label confirmed by the five majority voting or the reviewer that can be used for game-theoretic analysis or deep learning algorithms; (ii) how the changes affect the individual efficiency, or the person time needed for an individual labeler or reviewer to provide or check a label. In addition, we examine whether other desired properties of the data collection process, such as exhaustiveness, have been achieved.

To analyze efficiency, we first went through the person time data collected during VIOLA's development. Any changes made were deployed immediately to make faster progress. These person time data came from different videos and labelers. They inherently took different amounts of time to label, since the videos varied in their content. To mitigate the intrinsic heterogeneity, we divide the videos into four groups, $(0, 1)$, $[1, 2)$, $[2, 3)$, and $[3, +\infty)$, based on the average number of labels per frame, since it was an important indicator of the difficulty of labeling a video. There were other factors affecting the difficulty of labeling videos, so videos in the same group may still have had high variation. Because of this, we remove the top and bottom 5% of time per label entries.

Also due to these concerns, we collected additional person time data in a more controlled environment. We gave six unique videos that contained animals but no humans to the labelers to label. The labelers had not seen these videos previously. We distributed the work among the labelers so as to get one set of final labels for each video under each of the versions of VIOLA (as shown in Table 3.1). We asked the labelers to label for no more than 15 minutes on each video. To accommodate the labelers' schedules and coordinate their schedules to set up meetings, which are necessary for LabelDays and Tracking, we gave the labelers 2 to 4 days to label the videos under each version. As such, it was difficult to get multiple sets of labels for each video or get labels for more videos. Some labelers were not able to complete checking all of the frames in the video within 15 minutes, so we use the minimum checked frame among labelers for each video under each version, and analyze efficiency using person time data up until that frame only. Also, note that since some labelers were asked to label the same video multiple times under different versions, the labelers likely got faster as time went on. To mitigate these effects, we randomly ordered the five versions of VIOLA for them to label. The order is shown in Table 3.2.

We will proceed in this section by first focusing on the impact of the key change in the labeling framework from MajVote to LabelReview on the overall efficiency. We will then check each version of VIOLA to understand the impact of other changes. Because of the surprising results, we will

Table 3.2: Versions tested in the additional tests.

Version Number	1	2	3	4	5
Version Name	Basic	MultiBox	Review	LabelDays	Tracking
Framework Used	MajVote	MajVote	LabelReview	LabelReview	LabelReview
Test Order	Fourth	Third	First	Second	Fifth

particularly examine videos in which these features helped and in which they did not.

3.7.1 From MajVote to LabelReview

Fig. 3.8a and Fig. 3.8b show the comparison on overall efficiency between MajVote and LabelReview. The total person time per final label is lower on average when we use LabelReview, based on data collected through both the development process and additional tests. During the development process, there were only seven videos for which we got final labels from five full sets of labels using MajVote, two of which did not produce any consensus labels. There were more than 70 videos for which we got final labels through LabelReview. During the additional tests, we tested two versions using MajVote and three versions using LabelReview, which means the value of each bar is averaged over two or three samples, respectively. We exclude one sample for Video C where no consensus labels were achieved through MajVote. The LabelReview efficiency for Video D is 0.63 with a standard error of 0.09 but it is too small to appear in Fig. 3.8b.

In addition to having more labelers involved, one reason that MajVote leads to a higher person time per final label is the lack of consensus. Fig. 3.9 shows that there were large discrepancies in the number of labels between individual labelers, which led to fewer consensus labels (zero in Videos I and M).

Fig. 3.10 shows that MajVote leads to many fewer final labels than LabelReview for the videos in the additional tests. This indicates that using LabelReview can get us closer to the goal of exhaustively labeling all of the objects of interest when compared to MajVote.

3.7.2 Impact of Other Changes

In this section, we examine the individual efficiency and overall efficiency of each version of VIOLA to analyze the impact of every other change we made during the development of VIOLA. For individual efficiency, we calculate person time spent per label for each individual labeler or reviewer, regardless of whether that label has been confirmed to be a final label.

We first show results of individual efficiency based on person time data collected during the development process in Fig. 3.11. Person times per label for each video submission are colored to represent the group which is decided by the average number of labels per frame. Video submissions are reported by submission date since the date submitted indicates which version of the application was used for the video. The dates on which features were added, given in Table 3.1, are used to color the background of the plot. Finally, each submission is considered separately, to examine labeling or review efficiency only. Fig. 3.11 shows the person time per label for videos with low average number of labels per frame $(0 - 1)$ is higher than others for both labeling and reviewing. Fig. 3.12 shows the mean labeling and reviewing time per label within the timespan of each change during the development process.

We next examine the individual efficiency for labeling and reviewing in the additional tests (Fig. 3.13). The results of each test have been shown by video, since there were only five sets of labels in the tests with MajVote (Version 1-2) and only one set of labels in the tests with LabelReview (Version 3-5). The five sets of labels in the MajVote tests are averaged by video, and the standard error bars are included. Fig. 3.13 shows that each of the changes we made resulted in an improvement on the individual efficiency for some, but not all, of the videos.

Multiple Box Selection The feature of multiple box selection was added to improve the individual efficiency of labeling. Checking the first two groups in Fig. 3.12 and Fig. 3.13, we notice that surprisingly, this feature improves individual efficiency for some of the videos (e.g., Video F), but not all of the videos. One possible explanation is that in videos where there are many animals that do not move much over time, the changing position of the bounding boxes is mainly due to the movement of the camera. In this case, using multiple box selection and moving all of the bounding boxes in the same direction simultaneously is helpful. However, in other videos where there are only one or two animals in each frame, it may be faster to move the boxes separately, particularly if an animal moves.

Labeling Days Labeling days were introduced with the aim to increase the overall efficiency. Fig. 3.14 shows the average person time per final label has slightly reduced from Review to Label-Days during the additional tests, and the person time per final label has reduced for Videos A, C, and F. Fig. 3.14 also shows the number of final labels has remained the same on average. The results indicate that introducing labeling days may help improving the efficiency and exhaustiveness of labeling, at least for some more complex videos. Subjective feedback from the labelers also indicated that introducing labeling days made it easier for them to deal with ambiguous cases, when it is difficult to maintain consistency and accuracy despite the guidelines. However, Fig. 3.12, Fig. 3.13, and Fig. 3.14 show that introducing labeling days does not lead to an improvement on individual efficiency in all cases. It is possible that it increased the individual labeling time due to extra discussion, but it may have saved time during review. We plan to analyze the effects of labeling days in more detail in the future.

Tracking The tracking feature is the newest feature. We included it in the additional tests but it has not been deployed for the labelers to use. During the tests, we received positive feedback from labelers, particularly on videos in which animals were far apart and bright. In addition, the tracking feature was able to successfully track two animals in the first 10% of Video B, as shown in Fig. 3.7. Unexpectedly, the initial results from the additional tests do not show a positive effect on time per label or number of labels. We believe this is due to the fact that it does not find a brightness threshold automatically, and is likely to track the wrong object when multiple objects are within the same search region. We plan to continue developing this feature given its promise in the cases where animals are far apart and bright.

Summary This section thus shows that while some of our proposed improvements actually led to increased efficiency, particularly the switch from MajVote to LabelReview, in other cases (e.g., multiple box selection), surprisingly, it only increased efficiency in some videos. This result indicates that we must not simply add features on the intuition that they are bound to improve performance, as they may only be useful for certain videos.

3.8 Conclusions

In conclusion, we presented VIOLA, which provides a labeling and reviewing framework to gather labeled data from a small group of people in a secure manner, and a labeling interface with both general features for difficult video data, and specific features for our green security domain to track wildlife and humans. We analyzed the impact of the framework and the features on labeling efficiency, and found that some changes did not improve efficiency in general, but worked only in particular types of videos. We will now use the dataset to train deep neural networks to automatically detect wildlife and humans in real-time.

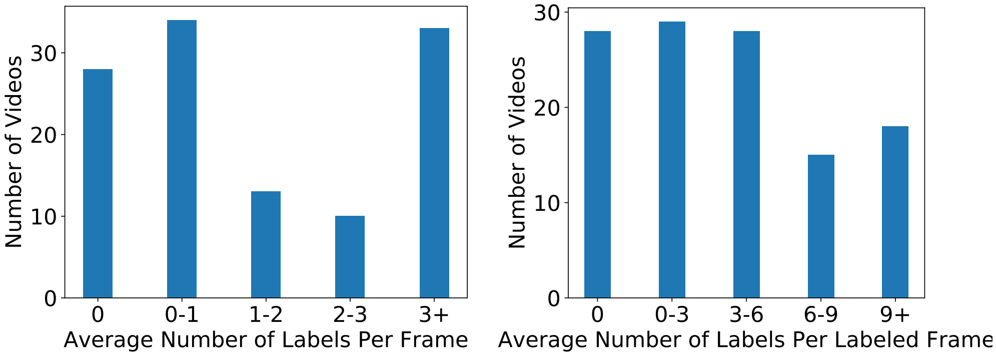
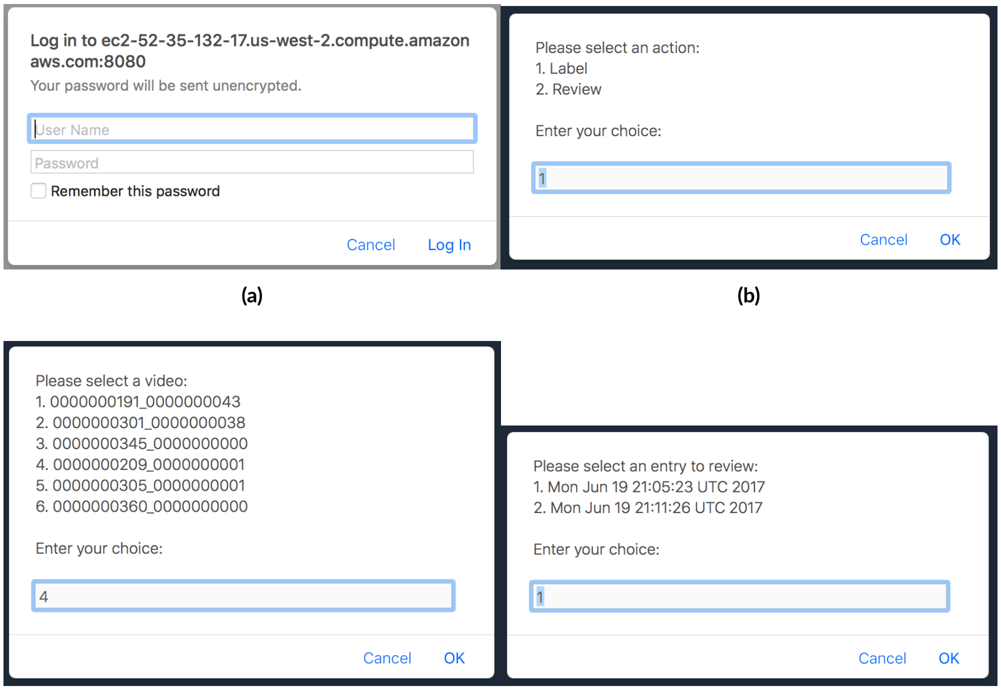


Figure 3.2: A histogram with the number of videos for average objects of interest per frame (left), and the average objects of interest per labeled frame (right).



Log in to ec2-52-35-132-17.us-west-2.compute.amazon aws.com:8080

Your password will be sent unencrypted.

(a)

Please select an action:
1. Label
2. Review

Enter your choice:

1

(b)

Please select a video:
1. 0000000191_0000000043
2. 0000000301_0000000038
3. 0000000345_0000000000
4. 0000000209_0000000001
5. 0000000305_0000000001
6. 0000000360_0000000000

Enter your choice:

4

(c)

Please select an entry to review:
1. Mon Jun 19 21:05:23 UTC 2017
2. Mon Jun 19 21:11:26 UTC 2017

Enter your choice:

1

(d)

Figure 3.3: The menus to begin labeling.

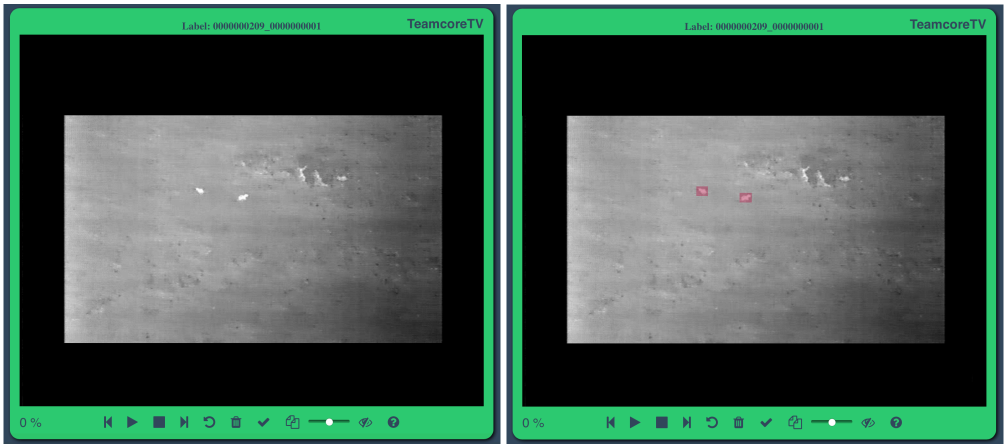
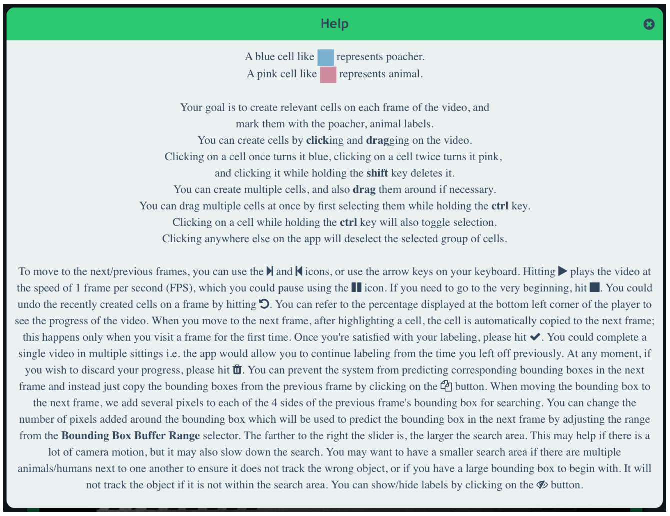


Figure 3.4: An example of a frame (left) and labeled frame (right) in a video. This is the next screen displayed after all of the menus, and allows the labeler to navigate through the video and manipulate or draw bounding boxes throughout.



A blue cell like 🟦 represents poacher.
A pink cell like 🟥 represents animal.

Your goal is to create relevant cells on each frame of the video, and
mark them with the poacher, animal labels.
You can create cells by **click**ing and **drag**ging on the video.
Clicking on a cell once turns it blue, clicking on a cell twice turns it pink,
and clicking it while holding the **shift** key deletes it.
You can create multiple cells, and also **drag** them around if necessary.
You can drag multiple cells at once by first selecting them while holding the **ctrl** key.
Clicking on a cell while holding the **ctrl** key will also toggle selection.
Clicking anywhere else on the app will deselect the selected group of cells.

To move to the next/previous frames, you can use the ⏭ and ⏮ icons, or use the arrow keys on your keyboard. Hitting ▶ plays the video at the speed of 1 frame per second (FPS), which you could pause using the ❚❚ icon. If you need to go to the very beginning, hit ■. You could undo the recently created cells on a frame by hitting ↺. You can refer to the percentage displayed at the bottom left corner of the player to see the progress of the video. When you move to the next frame, after highlighting a cell, the cell is automatically copied to the next frame; this happens only when you visit a frame for the first time. Once you're satisfied with your labeling, please hit ✔. You could complete a single video in multiple sittings i.e. the app would allow you to continue labeling from the time you left off previously. At any moment, if you wish to discard your progress, please hit 🗑. You can prevent the system from predicting corresponding bounding boxes in the next frame and instead just copy the bounding boxes from the previous frame by clicking on the ⧉ button. When moving the bounding box to the next frame, we add several pixels to each of the 4 sides of the previous frame's bounding box for searching. You can change the number of pixels added around the bounding box which will be used to predict the bounding box in the next frame by adjusting the range from the **Bounding Box Buffer Range** selector. The farther to the right the slider is, the larger the search area. This may help if there is a lot of camera motion, but it may also slow down the search. You may want to have a smaller search area if there are multiple animals/humans next to one another to ensure it does not track the wrong object, or if you have a large bounding box to begin with. It will not track the object if it is not within the search area. You can show/hide labels by clicking on the 👁 button.

Figure 3.5: Help screen detailing the controls of the application (? icon).

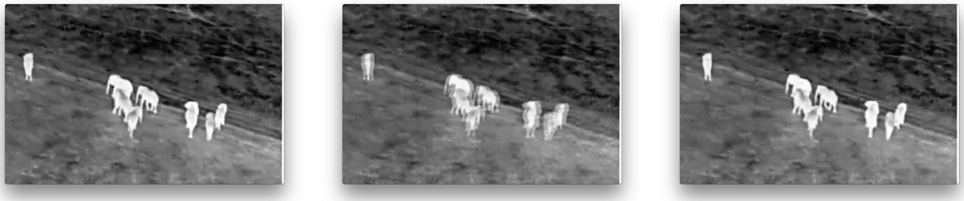
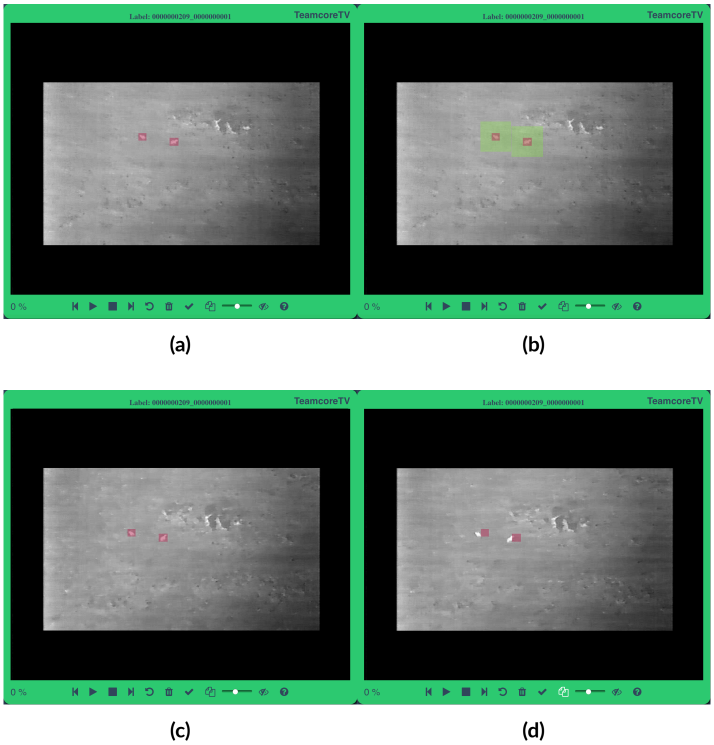


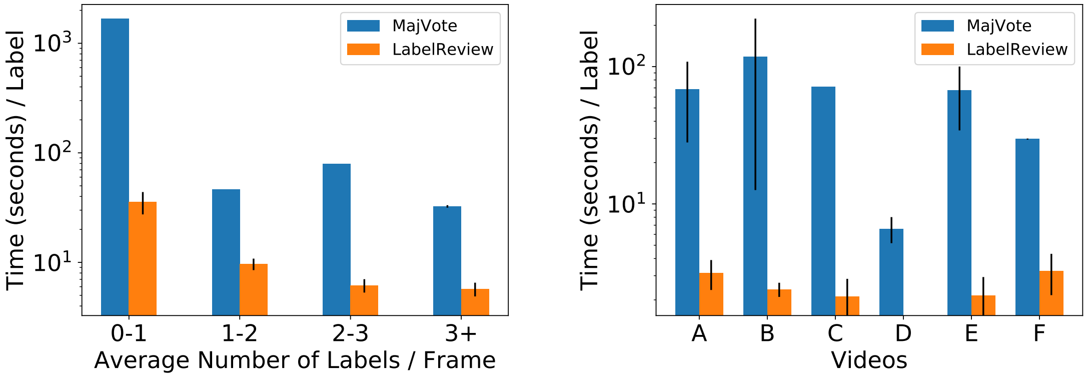
Figure 3.6: Three consecutive frames where the middle frame has ghosting. The middle frame is "in between" the left and right frames.


Label: 0000000209_0000000001
TeamcoreTV
0 %
Label: 0000000209_0000000001
TeamcoreTV
0 %
Label: 0000000209_0000000001
TeamcoreTV
0 %
Label: 0000000209_0000000001
TeamcoreTV
0 %

(a) (b)

(c) (d)

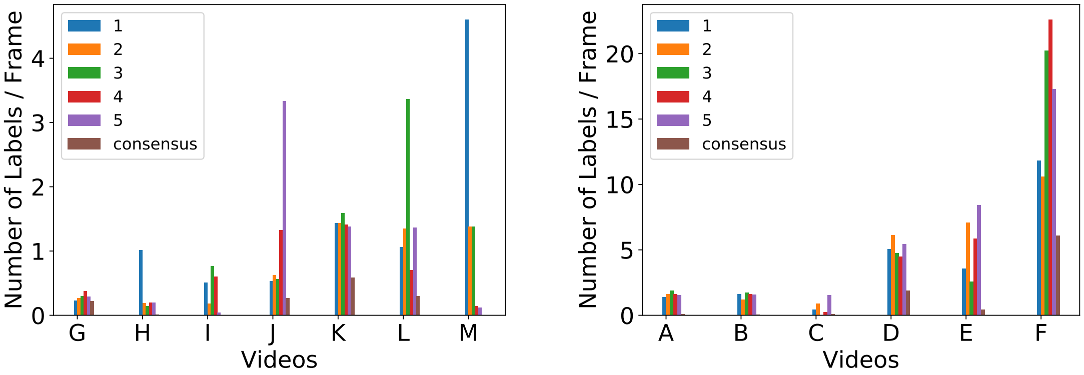
Figure 3.7: A sample labeling process.



(a) Data from development process.

(b) Data from additional tests.

Figure 3.8: Comparison of overall efficiency with different labeling frameworks.



(a) For the seven videos with five sets of labels during development process.

(b) For the six videos used in the additional tests under version Basic.

Figure 3.9: Number of labels per frame for individual labelers and for consensus.

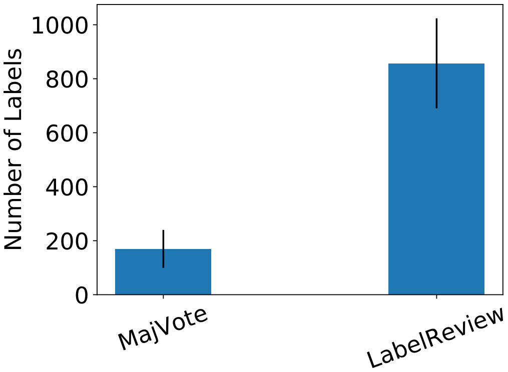


Figure 3.10: Number of final labels for MajVote and LabelReview in additional tests.

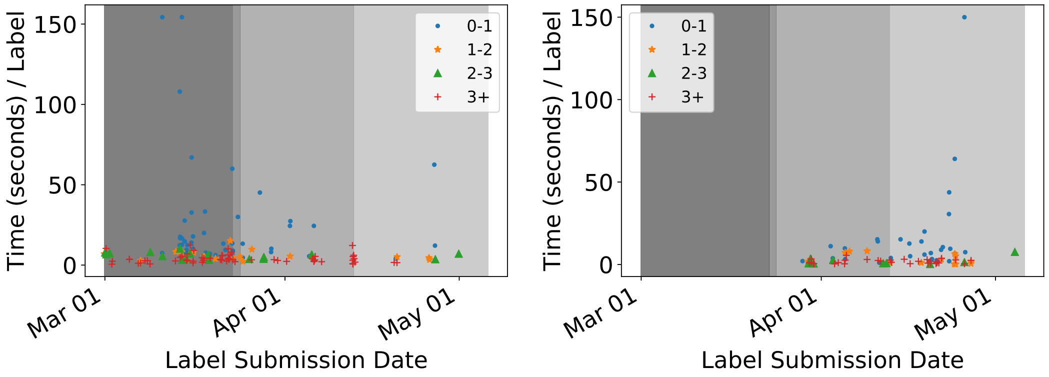


Figure 3.11: Individual efficiency for each submission of labeling (left) and review (right) with data collected during the development process.

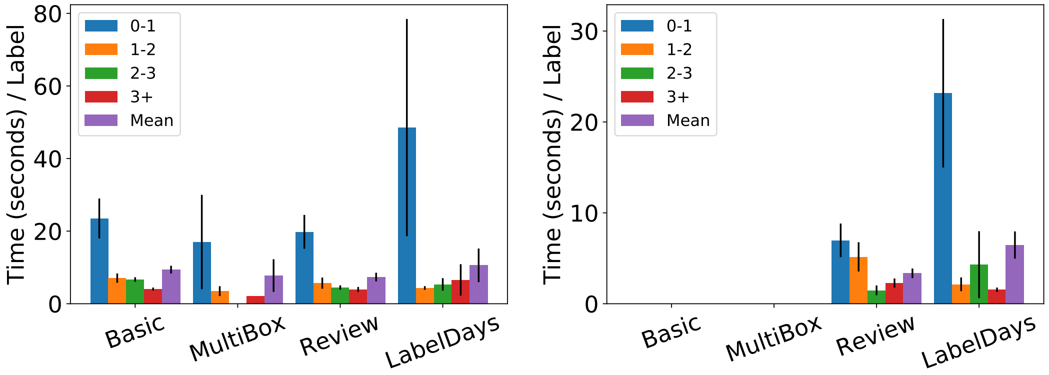


Figure 3.12: Average individual efficiency of labeling (left) and review (right) with data collected during the development process.

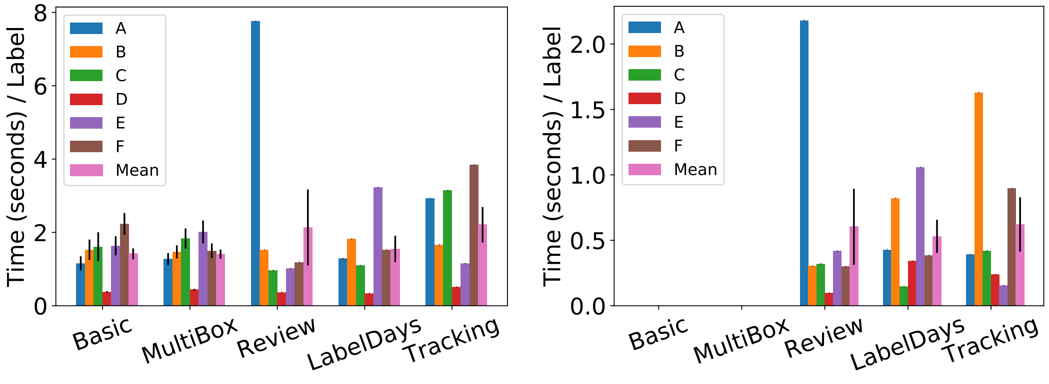


Figure 3.13: Individual efficiency for each submission and average efficiency of labeling (left) and review (right) with data collected from the additional tests.

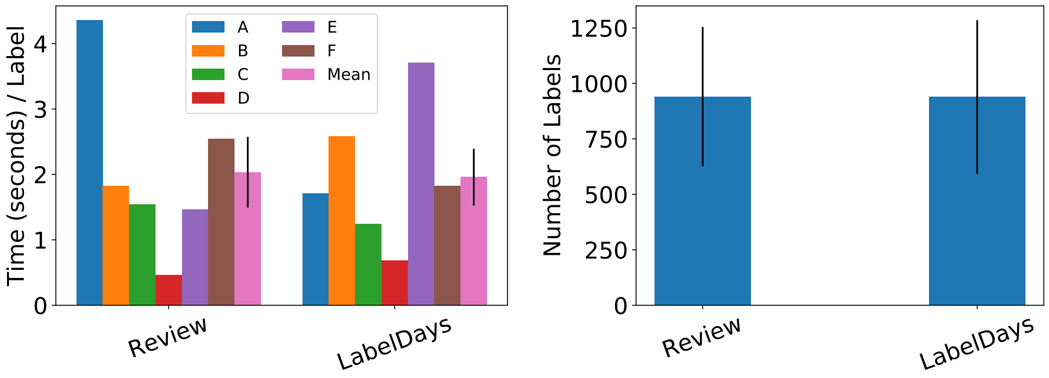


Figure 3.14: Overall efficiency (left) and number of final labels (right) with version Review and LabelDays during the additional tests.

4

Augmenting Conservation Drones with Automatic Detection in Near Real Time

4.1 Introduction

With elephant and rhino numbers dropping rapidly[112], it is imperative that we swiftly act before they are hunted to extinction. Multiple strategies exist to combat poaching, including park ranger

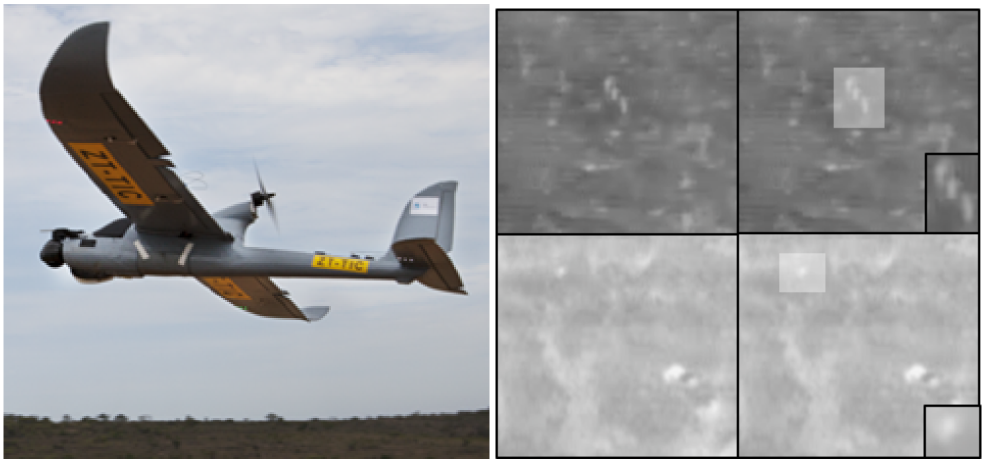


Figure 4.1: Example conservation drone and thermal frames from conservation drone, with white boxes surrounding humans.

patrols, and more recently, the use of conservation drones [137]. In particular, conservation drones equipped with long wave thermal infrared (hereafter referred to as thermal infrared) cameras can be used for nighttime surveillance to notify park rangers of poaching activity because there is increased poaching activity at night, and because animals and humans are warm and emit thermal infrared light even at night. However, the video stream from these conservation drones must be monitored at all times in order to notify park rangers of humans. Monitoring of streaming footage is an arduous task requiring human supervision throughout the night, and is also prone to systematic lapses in quality as human detection often degrades with fatigue [241]. Furthermore, as more drones are added to the system, more resources are required to monitor the additional videos.

Whereas previous work in AI has focused on game theory for patrol planning [324,308] and machine learning-based poaching prediction [104,78] to assist human patrollers in combating poaching, little effort has focused on decision aids to assist the conservation drone crew in detecting humans and animals automatically. Given the tedious work of monitoring conservation drone videos, such a decision aid is in high demand. It could help reduce the burden of the monitoring personnel and the probability of missing poachers by simply notifying personnel or park rangers of a detection.

In the future, the decision aid could also be integrated with existing tools that predict poaching activity and guide human patrols. For example, the system could scout ahead for humans to protect park rangers, monitor in other directions than human patrollers, or gather more information about the location of wildlife for better predictions. The integration would lead to a new generation of machine learning and game theoretic tools to guide rangers and conservation drones simultaneously.

In building this decision aid, there are several major challenges. First, automatic detection in thermal infrared videos captured aboard conservation drones is extremely difficult, because (i) the varying altitude of the conservation drone can lead to extremely small humans and animals, possibly less than 20 pixels in the images, (ii) the motion of the conservation drone makes stabilization, and consequently human and animal motion detection, difficult, and (iii) the thermal infrared sensor itself leads to lower resolution, single-band images, much different from typical RGB images. Second, we must provide notification in near real time so the conservation drone can immediately start following humans in order to provide park rangers with current locations. Real-time detection is an especially difficult challenge because we have limited computing power and Internet in the field.

In this paper, we present SPOT (Systematic POacher deTector), a novel AI-based application that addresses these issues and augments conservation drones with the ability to automatically detect humans and animals in near real time. In particular, SPOT consists of (i) offline training and (ii) online detection. During offline training, we treat each video frame as an image, and use a set of labeled training data collected for this application[39] to fine-tune a model which has shown success in detecting objects of interest in images, Faster RCNN. During online detection, the trained model is used to automatically detect humans and animals in new frames from a live video stream, *showing that modern computer vision techniques are capable of conquering difficulties that have not been addressed before*.

We also use a series of efficient processing techniques to improve the online detection speed of SPOT in the field. Online detection can be completed either on the cloud or on a local computer.

Therefore, we have experimented with several architectures that trade off between local and remote computers, depending on network strength. Finally, we evaluate SPOT on both historical videos and a real-world test run in the field by the end users, a conservation program called Air Shepherd[6]. The promising field test results have led to a plan for larger-scale deployment, and encourage its use in other surveillance domains.

4.2 Problem Domain and Current Practice

Conservation programs such as Air Shepherd[6] send crews to fly conservation drones (Fig. 4.1) in national parks in Africa, including Liwonde National Park in Malawi and Hwange National Park in Zimbabwe, in order to notify park rangers of poaching activity. Teams of people are required for conservation drone missions, including several conservation drone operators and personnel capable of repairing the conservation drones should anything happen. The conservation drone is a fixed-wing aircraft with a range of 50 km and a flight time of 5 hours with one battery. It carries a FLIR 640 Vue Pro thermal infrared camera. The conservation drone flight path is pre-programmed based on typical poaching hotspots or tips. While flying at night, the conservation drone operators monitor the live video stream, transmitted via radio waves, for any signs of humans. Should anyone be spotted, the team will manually take control to follow the suspects, notify nearby park rangers, who are sometimes on patrol or in a van with the team, and guide them to the humans.

However, as we already mentioned, monitoring these videos all night is difficult. Several example frames from thermal infrared videos are shown in Fig. 4.1, with objects of interest highlighted in transparent white boxes on the right. Notice that these frames are grayscale, with few pixels on objects of interest and many objects that look similar to those of interest. It is often difficult for humans to recognize objects in these videos because of this, leading to recognition errors and hours of tedious work. As such, there is great need for a tool that automatically detects humans and animals, the objects of interest in these videos for conservation. This tool should provide detections with as much accuracy as possible in near real time speeds on a laptop computer in the field with a potentially slow Internet connection.

There has been some effort towards automatic detection. EyeSpy[120], the application that is used in current practice, detects moving objects based on edge detection. When in use, it first asks the monitoring personnel to provide parameters such as various edge detection thresholds and sizes of humans in pixels. EyeSpy then requires information such as altitude and camera look angle throughout the flight to complete detection. Three limitations restrict the use of this tool as a result. First, EyeSpy relies heavily on a well-trained expert who can manually fine-tune the parameters based on the conservation drone and camera information. Novices are often unable to find the correct settings. Second, the parameters need to be compatible with flight altitude and camera look angle. To make this tool usable, the conservation drone crew either needs to restrict the way the conservation drone flies by keeping the flight altitude and camera look angle almost the same throughout the mission, or have the expert monitoring personnel manually adjust the parameters from time to time as the settings change. Third, this tool cannot differentiate between wildlife and humans, and thus cannot highlight the detection of humans to the monitoring personnel or the patrol team. We will examine this tool further in Evaluation.

4.3 Related Work and Design Choices

We arrive at the current framework of SPOT after several rounds of trials and errors. As humans and animals are typically warmer than other objects in the scene, and consequently brighter, we first consider automatic thresholding techniques such as Otsu thresholding[229]. However, other objects such as vegetation often have similar digital counts and lead to many false positives (Fig. 4.2(b)). Because humans and animals tend to move, we also consider motion using algorithms such as the

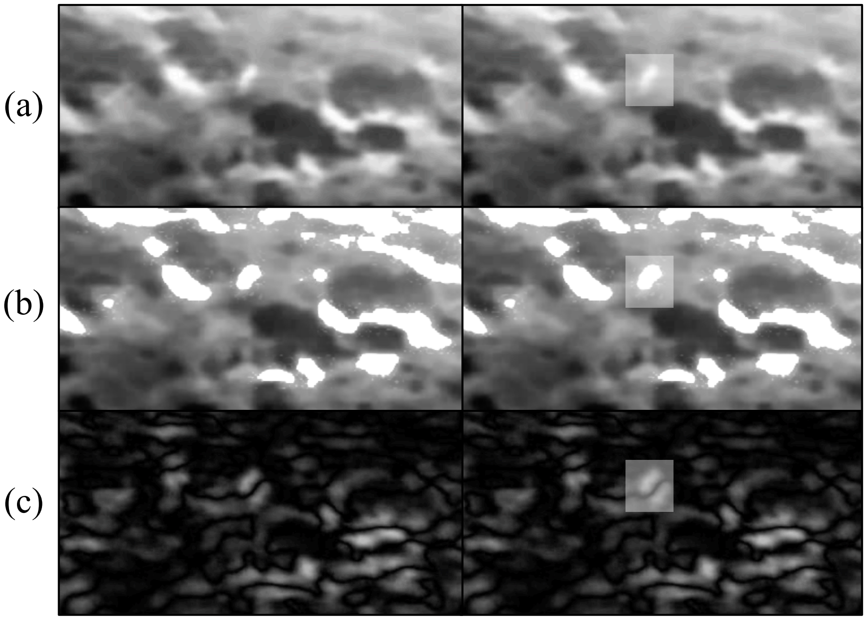


Figure 4.2: Traditional computer vision techniques. (a): original image, (b): thresholded, where white pixels are above the threshold, (c): stabilized frame difference. Original results (left), manually added transparent white boxes around true humans (right). These figures illustrate the difficulty these techniques face in locating humans.

Lucas-Kanade tracker for optical flow[186] and general correlation-based tracking[189]. Again, other objects such as vegetation look similar to the objects we want to track, which often leads to lost or incorrect tracks (Fig. 4.1). Assuming a planar surface, small moving objects can also be detected by a background subtraction method after applying video stabilization[230]. Motion is unfortunately detected incorrectly by this method in the case of complex terrain such as tall trees (Fig. 4.2(c)). More complex algorithms to track moving objects throughout videos rely on high resolution, visible spectrum videos or videos taken from a fixed camera[162,204].

Given the limitations of these traditional computer vision techniques and the great strides in object detection using convolutional neural networks, we turn to deep learning-based approaches. We treat each frame of the video as an image, and apply techniques to localize and classify the objects of interest in the images. Faster RCNN and YOLO[249,247] are two state-of-the-art algorithms suitable for this purpose. They both propose regions automatically for classification. Faster RCNN tends to have higher accuracy than YOLO, particularly for smaller objects, although YOLO tends to be faster[247]. A newer version, YOLOv2,[248], has improved performance over YOLO and could be used as an alternative to Faster RCNN. In this work, we focus on using Faster RCNN for detection.

Other emerging techniques such as deep learning-based optical flow or tracking[346,92] may fail due to drastic conservation drone motion and low resolution frames, and they do not classify the objects, only localize. Tubelets[147] propose bounding boxes over time, but are not yet performing in real time even on GPUs. Recently, there has also been some work on automatic wildlife detection and counting based on videos from conservation drones using other traditional computer vision or machine learning techniques, but they either rely on RGB images in high resolution[227] or do not consider real-time detection[302]. Due to the unique challenges of our problem, these techniques cannot be applied to detecting humans during flights at night.

4.4 SPOT

4.4.1 Overview

SPOT includes two main parts: (i) offline training and (ii) online detection (Fig. 4.3). In this section, we introduce both parts in detail, with an emphasis on the robust and faster processing techniques we use to improve the online detection efficiency and provide detections in near real time.

4.4.2 Offline Training

In our problem, detection means to localize the objects of interest in the scene, and classify them as humans or animals. We choose a state-of-the-art object detection algorithm, Faster RCNN, to serve our purpose. Faster RCNN is composed of a region proposal network (RPN) and Fast Region-based Convolutional Network method (Fast RCNN)[109], which is used to classify regions from the

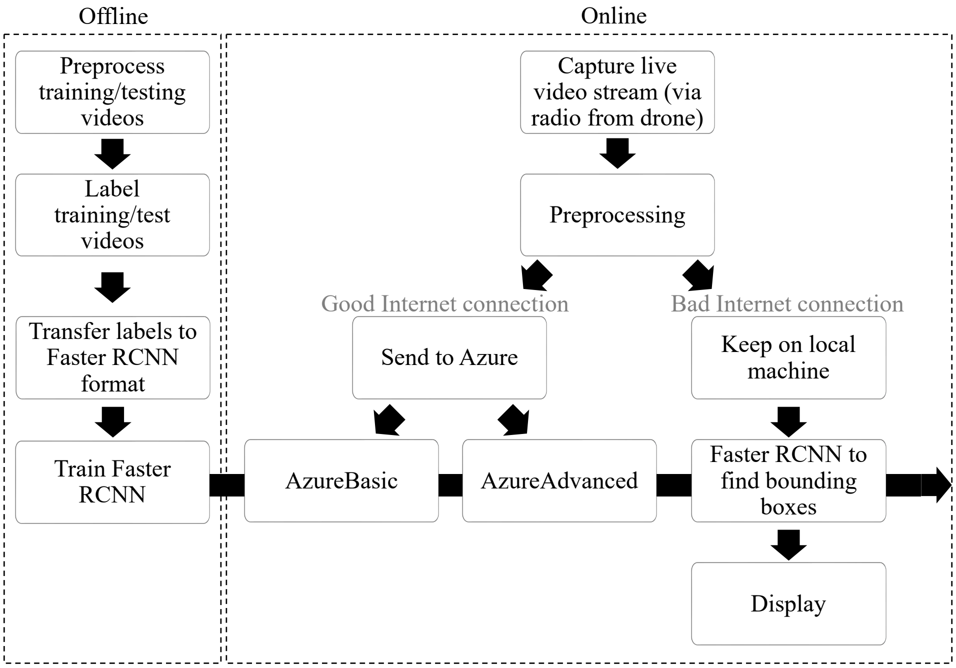
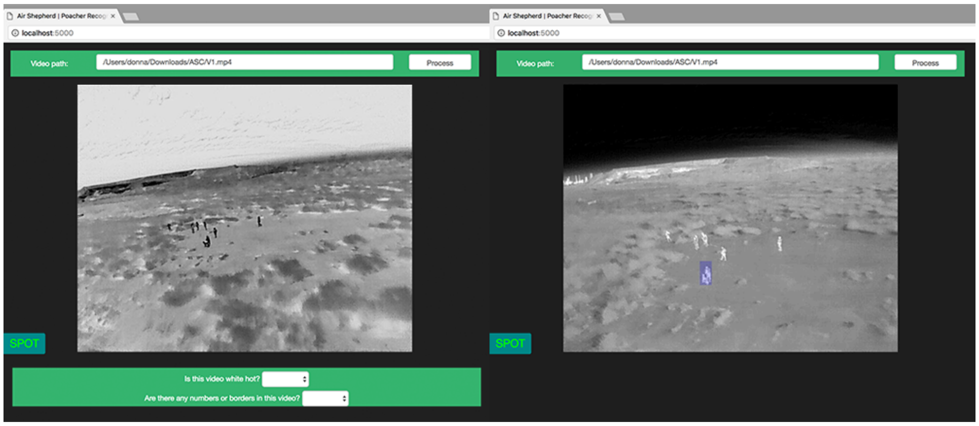

Offline
Online
Preprocess training/testing videos
Label training/test videos
Transfer labels to Faster RCNN format
Train Faster RCNN
Capture live video stream (via radio from drone)
Preprocessing
Good Internet connection
Bad Internet connection
Send to Azure
Keep on local machine
AzureBasic
AzureAdvanced
Faster RCNN to find bounding boxes
Display

Figure 4.3: SPOT Overview.

RPN, thereby giving us the location and class of objects. The RPN shares the convolutional layers of Fast RCNN, which is VGG-16[274] in our system.

To train the Faster RCNN model, we first initialize the VGG-16 network in the Faster RCNN model with pre-trained weights from ImageNet. Then, we use a set of videos in this application domain with annotated labels for each frame, collected using a framework described in[39]. A small team of students (not Amazon Mechanical Turk users in order to protect sensitive information such as flight locations and strategies) used this framework to label all frames in 70 videos containing animals and humans. Because consecutive frames are similar, we do not have enough heterogeneous data samples to train VGG-16 from scratch. This is the reason we start with pre-trained weights and fine-tune VGG-16 by treating each video frame as a separate image. Furthermore, we fine-tune different models for human and animal detection, so that depending on the mission type, whether



Air Shepherd | Poacher Recog

localhost:5000

Video path: /Users/donna/Downloads/ASC/V1.mp4

Is this video white hot?

Are there any numbers or borders in this video?

(a)

Air Shepherd | Poacher Recog

localhost:5000

Video path: /Users/donna/Downloads/ASC/V1.mp4

(b)

Figure 4.4: GUI created for SPOT for use in the field. 4.4a: inquiries about video, 4.4b: detection.

monitoring a park for poachers or counting animals, for example, the user may choose a model to provide better detection results. For the human-specific model, we fine-tuned using 4,183 frames, and for the animal-specific model, we used 18,480 frames, as we have more animal videos.

4.4.3 Online Detection

Preprocessing

Thermal infrared images can be "black-hot" or "white-hot", meaning warm objects are darker or lighter, respectively. During the online detection, we first ask the user if the video is white-hot, and if the answer is no, we will invert every frame we receive from the conservation drone. In addition, there is occasionally a border or text on the videos, consisting of date, flight altitude, and other meta-data. We ask users to provide the area of interest at the beginning and only display detections in this area of interest throughout the flight.

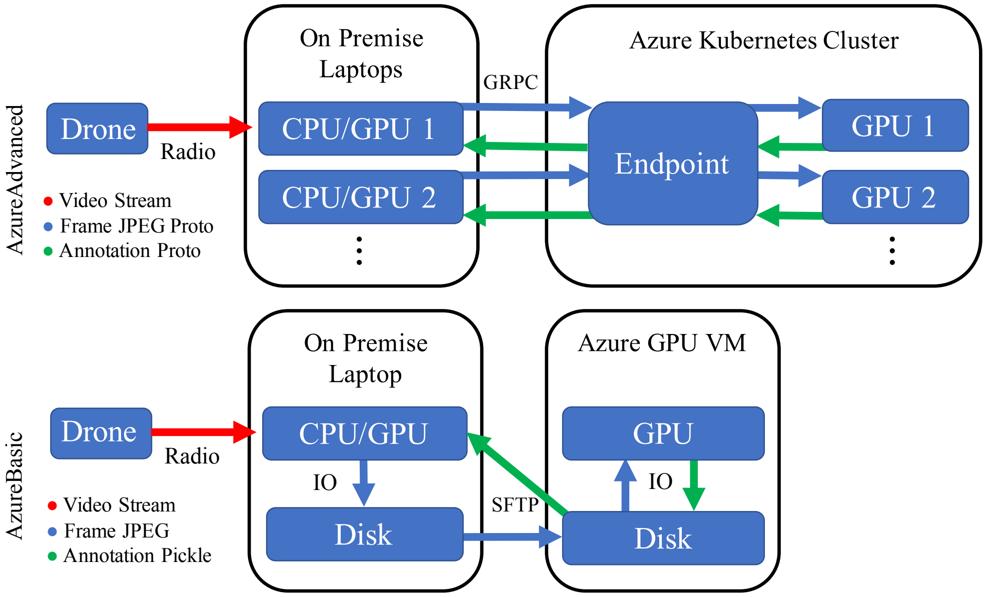


Figure 4.5: AzureBasic and AzureAdvanced overview.

Detection

We treat each frame of the video stream as an image and input it to Faster RCNN. The trained model computes regions and classes associated with each region.

User Interface

Fig. 4.4 shows the user interface of SPOT for online detection. A file can be selected for detection, or a live stream video. In Fig. 4.4a, we gather preprocessing information about the video, and then begin detection in Fig. 4.4b.

Architectures and Efficiency

Faster RCNN runs at 5 frames per second (fps) on a K40 GPU[249]. Efficiency and computation speed are paramount for similar performance in the field where there may be limited computing power, especially since videos are 25 fps. We therefore examine different Microsoft Azure architectures (Fig. 4.5), and discuss techniques to improve performance in the field and trade off between local and remote compute.

The first and simplest cloud architecture we investigate, which we will refer to as AzureBasic, is an NC-6 Series Virtual Machine (VM) with a Tesla K80 NVIDIA GPU hosted on Microsoft Azure. We simply transfer frames from the local laptop to this VM using Paramiko, a Python SFTP client. Once frames are transferred to the remote machine, we detect objects in the frame using our stored, fine-tuned Faster RCNN model in a running Python instance on the remote machine. We then display the annotated frame locally using X forwarding. For the purposes of testing, we send frames in batches, and we use Paramiko to transfer annotated frames instead of displaying. Speed could be improved by transferring annotations instead of annotated frames.

Although AzureBasic allows us to improve our throughput through cloud GPU acceleration over a CPU laptop, it is limited to a single Azure GPU VM and a single local laptop linked together by SFTP. To scale out SPOT, we utilize Tensorflow Serving, a framework for efficiently operationalizing trained Tensorflow computation graphs. Tensorflow Serving provides a way to evaluate Faster RCNN without the overhead of a running Python instance and file IO from SFTP. Furthermore, Tensorflow Serving communicates through Protocol Buffers, a flexible and efficient data representation language that significantly reduces the size of large tensors. For serving scenarios with large requests and responses, such as video processing, this reduces network communication and improves performance on slow networks. Tensorflow Serving also supplies tools for creating multi-threaded clients. We use four threads for our testing. Like AzureBasic, we also process images in batches to ensure that there is no downtime between uploading frames and downloading the results. Finally, we use azure-engine to create a cluster of NC-6 series GPU VMs managed with Kubernetes, a fault tolerant load balancer for scalable cloud-based services. This keeps the latency of SPOT low in potential compute intensive multi-conservation drone scenarios. It also provides a single REST end-point so the client code can use a single web URL for sending images regardless of the number of machines in the cluster. We deploy on a GPU-enabled docker image with Tensorflow Serving, and add tools for convenient re-deployment of models hosted on Azure Blob Storage. We refer to this architecture as AzureAdvanced.

4.5 Evaluation

To provide a working prototype system, SPOT needs to meet two major criteria: (i) detection accuracy and (ii) efficiency. Detection accuracy is most important for poacher identification, particularly to make sure we have few false negatives and false positives. Speed is critical to being able to quickly notify monitoring personnel and the ranger team. In this section, we evaluate SPOT in the lab using six historical videos, consisting of 15,403 frames in total, as test video streams. We will first evaluate the performance of the object detection, and then the efficiency, where we compare the different methods discussed in earlier sections.

EyeSpy[120], the application that is used in current practice, requires users to tune eight parameters to correctly identify objects of interest, plus six flight metadata parameters such as altitude and camera angle. Because of so many parameters, it is often difficult to successfully tune all of these parameters as a novice. On the other hand, our application does not require the user to fine-tune any parameters – it can be used as is. We therefore consider EyeSpy as used by a novice (ESN). Of our six test videos, only the three animal videos have average flight metadata records (i.e., not flight metadata per frame). For analysis of ESN, we use flight metadata parameters if present, and make educated guesses for altitude if not, because this is the baseline only. Otherwise, we utilize default values for all parameters. We also include results from EyeSpy as used by an expert (ESE). These parameters are adjusted by our collaborators at Air Shepherd who created EyeSpy. We do not make educated guesses for ESE because a lack of exact parameters could drastically reduce performance of

Table 4.1: Precision-Recall for SPOT and EyeSpy Novice (ESN) for animals.

	Precision		Recall	
Video	SPOT	ESN	SPOT	ESN
SA	0.5729	0.1536	0.0025	0.0072
MA	0.5544	0.0032	0.0131	0.0058
LA	0.5584	0.0235	0.2293	0.0694

Table 4.2: Precision-Recall for SPOT and EyeSpy Novice (ESN) for humans.

	Precision		Recall	
Video	SPOT	ESN	SPOT	ESN
SH	0	0.00003	0	0.0007
MH	0.0995	0.0004	0.0073	0.0009
LH	0.3977	0.0052	0.0188	0.0159

EyeSpy, which would not be a fair comparison. We record the output from EyeSpy, which is a video with red outlines around objects of interest, and place bounding boxes around any red outlines obtained. We then use an IoU threshold of 0.5 as is typical in[249]. Finally, we choose a low confidence threshold for SPOT because missing a human detection is extremely undesirable, and we report the precision and recall.

We compare the performance of SPOT and ESN on videos containing animals or humans with labels of small, medium, or large average sizes in Tables 4.1 and 4.2. We also compare the performance of SPOT and ESE in Table 4.3. We perform better than the novice in both precision and recall for medium- and large-sized humans and animals. We also perform better than the expert for large-sized animals, and comparably for small- and medium-sized animals. Because we perform better than ESN and similarly to ESE, we thus reduce significant burden. For small humans, which is a challenging task for object detection in general, both tools perform poorly, with EyeSpy being able to identify a small number of humans correctly. Small animals also prove to be a challenge for SPOT. To improve performance for small objects in the future, we expect pooling the results of video frames and incorporating motion will be beneficial.

Table 4.3: Precision-Recall for SPOT and EyeSpy Expert (ESE) for animals.

	Precision		Recall	
Video	SPOT	ESE	SPOT	ESE
SA	0.5729	0.6667	0.0025	0.0062
MA	0.5544	0.0713	0.0131	0.0162
LA	0.5584	0.0433	0.2293	0.0832

Table 4.4: Timing Results for CPU, AzureAdvanced (AA), AzureBasic (AB), and GPU.

	# GPUs	Network	s/img
CPU	0	-	10.4354
AB	1	fast	0.5785
AB	1	slow	2.2783
GPU	1	-	0.3870
AA	2	fast	0.3484
AA	2	slow	0.4858

Next, we evaluate efficiency by comparing CPU performance to the initial Azure system, to the improved Azure system, and finally to the single GPU performance. The GPU laptop is a CUK MSI GE62 Apache Pro, with Intel Skylake i7-6700HQ, 32GB RAM, and the NVIDIA GTX 960M with 2GB RAM. It is deployed in the field. The CPU laptop has an Intel i5-3230M CPU at 2.60GHz. In order to compare the Azure systems, we time how long it takes from the frame being sent to Azure, to the prediction, to the return back to the local machine, and finally to reading the final image back into memory. We conducted these tests in two different networking environments: 533.20 Mbps upload and 812.14 Mbps download, which we will call “fast”, and 5.33 Mbps upload and 5.29 Mbps download, which we will call “slow”. We repeat the experiment for several images and show the final time per image in Table 4.4. The results show that both AzureAdvanced and the GPU laptop perform detection almost 100 times faster than the CPU laptop, and AzureAdvanced drastically improves over AzureBasic when a slower network is present. Therefore, we can achieve detection in near real time.

Table 4.5: Precision-Recall for SPOT, EyeSpy Novice (ESN), and EyeSpy Expert (ESE) for humans in test video.

Precision			Recall		
SPOT	ESN	ESE	SPOT	ESN	ESE
0.4235	0.0024	0.0573	0.3697	0.0432	0.2836

4.6 Implementation in the Field

We also evaluate the in-field performance of SPOT. So far, these tests have been run by Air Shepherd at a testing site in South Africa, where training exercises take place. Fig. 4.6 shows a screenshot from a 30 minute test of AzureBasic at the site. For a full video, sped up 20 times, please visit http://bit.ly/SPOTVideo. Precision and recall results are shown for this in Table 4.5, which shows that SPOT performs better than both ESN and ESE. Our collaborators at Air Shepherd reported that SPOT performed human detection well during this test flight, and was so promising that they want to move forward with further development and deployment in Botswana. They also showed excitement because SPOT requires no tuning from the user. Although the network connection was poor for some of the flight and caused detection to occur slowly, especially because AzureBasic was used, AzureAdvanced will perform better in these situations, and the GPU laptop can now provide consistent detection speeds with slow networks, which our collaborators found encouraging as well. With the promising results from the field test, a wider deployment is being planned.

4.7 Lessons Learned and Conclusion

In conclusion, we developed a system, SPOT, to automatically detect humans as well as animals in thermal infrared conservation drone videos taken at night in near real time, which shows that modern computer vision techniques are capable of conquering difficulties that have not been addressed before. This system works in varying situations and does not require the users to adjust any parameters when they use it. Thus, it is easily accessible to non-expert users. Furthermore, the system can

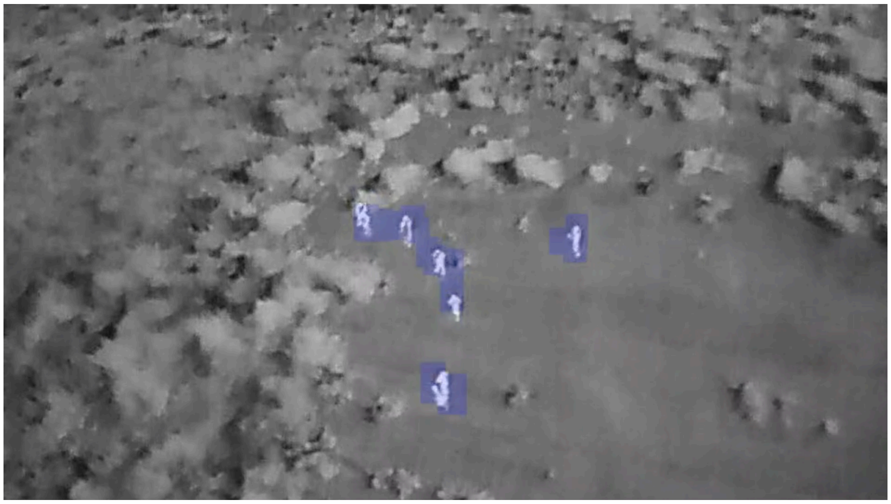


Figure 4.6: A screenshot of the field test environment with annotated figures.

detect humans in near real time with either good or bad network connectivity. The system has been tested in the field, and will be deployed.

5

AirSim-W: A Simulation Environment for Conservation Drones

5.1 Introduction

Wildlife conservation is one of the most important sustainability goals today, and innovations in artificial intelligence are uniquely suited to advancing it. When it comes to wildlife poaching in particular, artificial intelligence has already played an important mitigating role. In order to maximize the protection of national parks and conservation areas, it has been used to assist park rangers in planning their patrols to find poachers and snares, both in predicting future poaching incidents[104,78] and creating strategies to detect poaching or signs of poaching activity[93,324]. Recent advances in conservation drone technology have made conservation drones viable tools to aid in park ranger patrols. Conservation drones can play a role in patrolling by deterring poaching through the use of signaling[327], serving as a lookout for park rangers, or even acting as a separate patroller when equipped with the ability to automatically detect animals and humans in conservation drone videos.

The ability to detect animals and humans in conservation drone videos, particularly thermal infrared videos, is an active area of research due to the small size of humans and animals in conservation drone videos, the conservation drone motion, and the low-resolution, single-band nature of thermal infrared videos. In our previous work, a dataset of 70 historical thermal infrared videos was labeled[39]. These videos were collected by Air Shepherd between 2015 and 2017 during flights which typically occur at night based on pre-programmed paths. Flights go on for about 8 hours per night, with individual flights that are 2 hours long due to battery life. When objects of interest are observed on these flights, the conservation drones are flown manually in order to follow the objects of interest. Often, however, videos do not have many objects of interest, or it is difficult to identify objects of interest in the videos as human observers. This means that videos had to be checked for content first before labeling, which added additional time to the process. In total, **this labeling process took approximately 800 hours over the course of 6 months to complete**, and produced 39,380 labeled frames and approximately 180,000 individual human and animal labels on those frames. At a rate of $11 per hour, this cost about $8,800 for labeling alone, plus flying costs between 2015 and 2017. Together, the time and money associated with labeling make it extremely difficult to collect large labeled datasets like this.

Once the 70 videos had been labeled, individual frames were used to train Faster RCNN[249] for animal and human detection, which was part of a larger system called SPOT[38]. Training was completed on 22,663 total frames, with 18,480 total frames for the animal model and 4,183 frames for the human model. Note that these models each detect both animals and humans, but due to the random sampling, the animal model performed better at detecting animals than other tested models, and the human model performed better at detecting humans than the other tested models. SPOT performed better than the existing tool used by Air Shepherd.

Although SPOT is immediately useful to park rangers in the field as a decision aid, park rangers or others hired to monitor the videos are still required to confirm human detections made by SPOT and then manually fly the conservation drone to follow the human. In order to improve detection performance, more labeled training data is needed. Additionally, to further relieve the burden on rangers, we would like to allow for autonomous flight to follow planned patrol routes, deviate from the plans as needed to further investigate possible detections, and automatically follow detected humans. However, testing of autonomous flight in the field could be costly, as mistakes could lead to poached animals. Existing work does not address these unique challenges, so we propose a new method based on simulation of the domain environment. This allows us to augment our dataset of labeled thermal infrared videos efficiently, and to provide a testing environment for future autonomous flight and other costly experiments in the domain of wildlife conservation, such as patrol planning.

To build a simulation with these features, we use Unreal Engine and AirSim[267]. Unreal Engine is a game engine where various environments and characters can be created, and AirSim is a simulator for drones and cars built on Unreal Engine. AirSim supports hardware-in-the-loop (e.g., Xbox controller) or a Python API for moving through the Unreal Engine environments, such as cities, neighborhoods, and mountains. AirSim specifically consists of a vehicle model for the conservation drone, which is modeled as a quadrotor, an environment model, made up of gravity, magnetic field, and air pressure and density models, a physics engine for the linear and angular drag, accelerations, and collisions, and finally a sensor model for the barometer, gyroscope and accelerometer, magnetometer, and GPS. The models are created such that real-time flights are possible. As a result, scene, segmentation, and depth images can be collected during flights or drives through the environments, which allows artificial intelligence researchers to experiment with deep learning, computer vision, and reinforcement learning algorithms for autonomous vehicles.

In this work, we present AirSim-W, which includes the (i) creation of an African savanna environment in Unreal Engine, (ii) expansion of the current RGB version of AirSim to include a thermal infrared model based on physics, (iii) expansions to follow objects of interest or fly in zig-zag patterns to generate simulated training data, and (iv) demonstrated detection improvement using simulated data generated by AirSim-W. With these contributions, AirSim-W will be directly used for wildlife conservation research, especially for the challenges of human and animal detection in conservation drone videos and patrol planning for conservation drones and foot patrols.

5.2 Related Work

First of all, the main problem of interest is to utilize simulation for wildlife conservation. For the problem of automatic detection of wildlife and humans in conservation drone videos, in addition to SPOT[38], there has also been some work on wildlife counting based on videos from conservation drones using primarily traditional computer vision or machine learning techniques, including[227] and[302]. They either rely on RGB images in high resolution or do not consider real-time detection, and SPOT has shown improvement over a traditional computer vision result in near real time.

To improve on these results, we now examine data augmentation. Performance is often improved by increasing the amount of data used during training. For example, to train AlexNet[163], simple data augmentation involving cropping, translation, and horizontal reflections was utilized to increase the size of the training dataset by a factor of 2048, which helped reduce overfitting. They further augmented the dataset using PCAs to perturb digital counts. More recently, deep learning models such as generative adversarial networks (GANs) and recurrent neural networks (RNNs) have shown great promise in the realm of data augmentation[243,116,134,46,294]. In[243], deep convolutional GANs (DCGANs) are used to augment datasets and even draw certain objects, such as a bedroom. Style transfer and image-to-image translation are other areas being considered for data augmentation[185,345]. These could be used to take many images of horses and convert them to zebras, or convert images taken in daylight to nighttime images, all of which may help with a specific computer vision task. However, these methods (i) do not account for thermal infrared imagery, and (ii) do not consider the physical processes that are involved in image capture, such as movement of the image capture platform.

Further data augmentation is possible using simulation from computer graphics. There are many examples of environments that have been built using rendering tools such as Unity[257] and Unreal Engine[267]. Digital Imaging and Remote Sensing Image Generation (DIRSIG)[133] is another example, where facetized surface models can be generated using AutoCAD, 3ds Max, Rhinoceros, Blender3D, or SketchUp, for example. Some environments exist with physics engines that allow for testing robotics systems within the environment, such as autonomous cars or drones. There are many of these environments, but we will only mention AirSim and Gazebo[267,157]. In any case, datasets can then be generated using these environments. For example, SYNTHIA[257] is a dataset generated by capturing images in a city environment in Unity, and has shown improvement in semantic segmentation. GANs have been used to give simulated data like SYNTHIA a more realistic look and to further improve semantic segmentation[272,342]. Few models examine the thermal domain, except DIRSIG, which uses a full radiometric model for thermal simulation, and[88], which uses simple 3D CAD models of solitary objects and a basic radiometric model.

Specifically in domains where little training data exists, undertaking the task of labeling large amounts of data can be time consuming and tedious, and data augmentation may be a necessity. For example, in the self-driving car domain,[141] trains a model using only simulated data to improve over the same model trained entirely on real data, while testing on real data. Video games such as GTA V can also be used to collect eye movement data for driving[24]. Another example in which there is a specific domain that may not have enough data in existing datasets is[277], where a mapping from simulated data from Unity to more realistic simulated data is learned and applied, and the results are used to train reactive obstacle avoidance and semantic segmentation neural networks.

For the purposes of our wildlife conservation domain, we require (i) data augmentation capabilities for computer vision tasks and (ii) a full simulation environment for future development in ranger and conservation drone patrol planning and autonomous conservation drone flights for conservation purposes. As already mentioned, GAN models have shown promise in data augmentation, but they do not account for thermal infrared imagery and the physics behind image capture. Of the simulation environments mentioned, although all promising for data augmentation, only AirSim and Gazebo allow for future conservation patrol and autonomous flight testing.

In this paper, we seek to build a tool suitable for the wildlife conservation domain. We will utilize AirSim due to the ability to use Unreal Engine as the underlying rendering tool. We will generate our environment, which will be an African savanna, in Unreal Engine. This will allow us to capture images in real time with AirSim[267], and to control actual flight parameters for image capture. We will also create a basic model of the physical characteristics of thermal infrared cameras to expand the performance of Unreal Engine and AirSim for training data generation in the human and animal detection domain. Finally, we will utilize our novel simulation technique in the area of wildlife conservation in particular, and it can be downloaded and easily used here: https://github.com/Microsoft/AirSim.

5.3 African Savanna Environment

To effectively run simulation of thermal infrared imagery capture, we needed to build out an environment that was similar to biomes found in the central African savanna, when viewed through imagery captured at an altitude from 200 to 400 feet (61 to 122 m) above ground level (AGL). We used web-sourced target images and Google Earth to visualize the environment in several national parks where Air Shepherd has flown previously. Visual targets varied from wide-open savanna plains to dense forest, and flatland to craggy canyons. Because of this large range, we chose to develop a representative biome rather than a facsimile of an existing location. Key features were wide-open space, dense forest, a mid-density area, a water feature, road access, and humans and appropriate animals.

We first included the correct plants, animals, and humans. Flora in the area generally consists of baobab, acacia, and hookthorn trees, as well as brush and grass. We were able to find accurate vegetation models for each of the tree types from an existing 3D model vendor, SpeedTree. We were also able to find a variety of pre-animated and rigged animals including elephants, rhinoceroses, hippopotamuses, zebras, lions, and crocodiles in the Unreal Engine Marketplace. Animals can also be found at TurboSquid, another 3D model vendor. Note that while we have not seen hippopotamuses or crocodiles in real data, we are able to model them in this simulated environment, allowing us to train on features which are lacking in our dataset. This is extremely useful as it allows us to address issues such as missing data or class imbalances in data, which is another benefit of using simulation. Our three human characters were the only assets that were custom, and were created with Autodesk, leveraging animation created from a motion capture suit to give a realistic walking motion.

Then, the general flow for the environment creation follows typical game environment workflow. We created the one square mile flat terrain, then sculpted in hills and depressions for water, with the water in the center of the map. The Unreal Engine scale unit is 1 cm, so we started with a rectangular polygon of 6 feet in length to appropriately scale people and animals in the scene. Following this, we created spline-based movement of the actors before starting the scene dressing. The Path Follow plug-in, which can be found on the Unreal Engine Marketplace, was used to create the actor movement as it provided a better movement capability than the native UE4 spline-based movement.

We next started dressing the scene. A water plane was added and adjusted for the desired water level. Vegetation was added with the native paintbrush capability using various densities to reflect dense, mid, and sparse areas, and was repeated for each of the vegetation types. Instead of painting performance-reducing grass across the entire scene, textures were created to reflect the look we desired for improved performance during real-time video capture. A dirt road was cut into the scene and textured appropriately, and two vehicles were sourced to add to the scene.

The scene reflects three generic areas of vegetation density to support imagery targets across all three areas with three sets of humans added to the scene. A set of humans consists of three individual characters with each set following a spline in a large loop. We intersected the human loops with elephants on spline loops to capture images of both humans and animals together. Additionally, zebras were scattered across the environment and animals were clustered around the watering hole.

Overall, the Africa environment was created in approximately 3 working weeks with an artist and part-time developer, totaling approximately $5000 and about 180 hours. The bulk of the time spent on this scene was the terrain, watering hole, vegetation, and design of the NPC movement, with a lesser amount of time on creating the animal and human spline movement. Several example images from the environment are shown in Fig. 5.1.

Should these costs be unmanageable to those in the conservation domain when considering environments other than an African savanna, transfer learning is a low-cost possibility to consider in the future, especially because the Africa environment is being made freely available through Microsoft

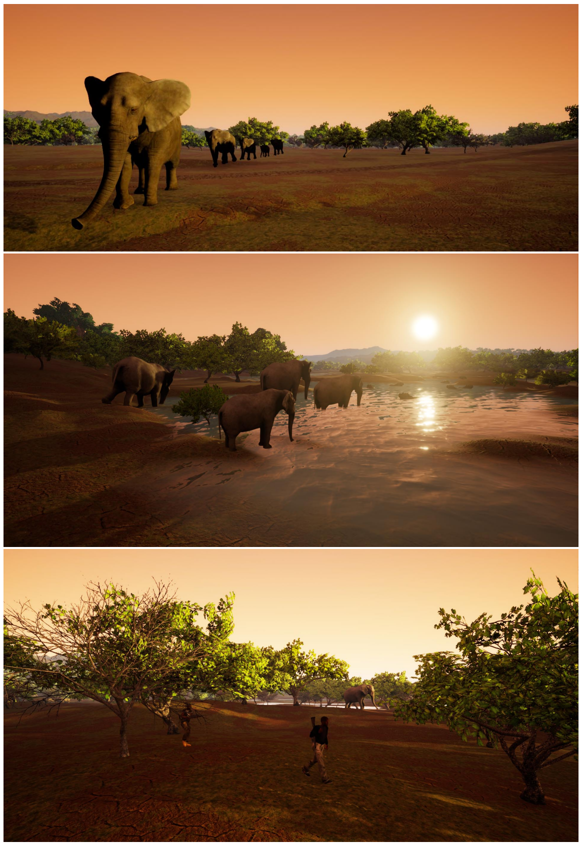


Figure 5.1: Example still images from the Africa environment.

AirSim (https://github.com/Microsoft/AirSim/releases). In addition, many of the assets used in the Africa environment came from the Unreal Engine Marketplace. There are likely environments, animals, and plants from other regions that could be simply bought and used directly. Together, these facts make creating an environment other than an African savanna for other domains possible at a relatively low cost.

5.4 Expansion from RGB to Thermal Infrared

5.4.1 Physical Modeling Assumptions

Although the African savanna environment is already useful by itself, we must expand it to include thermal infrared imagery in order to augment our dataset for detecting animals and humans in thermal infrared imagery. Simulated RGB imagery alone is not useful because flights are done at night, when RGB imagery is not available. Additionally, we pre-train Faster RCNN using ImageNet, which is a database including millions of RGB images that can be used by the network to understand edges and shapes before learning the specific thermal infrared image domain.

In order to simulate thermal infrared imagery from the RGB imagery in AirSim, particularly the resulting segmentation map, we will rely on physical modeling. Due to the large number of interactions between photons and objects in or near the scene, modeling light can become extremely complicated. In the thermal domain at night, for example, thermal light reaching the camera on a conservation drone could come from several different sources: (i) atmosphere at some temperature emitting thermal infrared photons directly into the sensor, (ii) atmosphere at some temperature emitting thermal infrared photons that hit the ground and are reflected by the target into the sensor, (iii) thermal infrared photons emitted directly from the target into the sensor (this can be modeled using Planck's Law[263]), and (iv) thermal infrared photons emitted by nearby objects that are then reflected by the target into the sensor. These different contributions are called upwelled, downwelled, direct, and background radiance, respectively[263]. In addition to the atmosphere contributing photons directly to the signal, it can also play a role whenever photons travel from the target to the sensor. Depending on whether it is humid, cloudy, rainy, etc., this role can be larger or smaller, and is often modeled by radiative transfer models such as MODTRAN[29]. Other effects on the signal include the uniformity with which the objects of interest emit light (e.g., whether or not they are Lambertian), camera spectral response, and camera sensor noise, especially non-uniformity correction in microbolometers.

Because all of this involves a significant amount of modeling of complex physical phenomena, we will make simplifying assumptions to create a simplistic physical model of the thermal infrared image that would result from objects in the African savanna at certain temperatures. First, upwelled radiance and downwelled radiance are negligible with a clear, dry, cool atmosphere. Most of the year this would hold true in Africa, except during rainy season in the summer, when flights are not likely to take place anyway. A clear, dry, cool atmosphere also has negligible effects on transmission. Background radiance is negligible in cases of mostly flat terrain, which generally applies in a savanna. This means that the dominant contribution is direct, so we do not consider the contributions of the atmosphere to the signal, nor do we consider the transmission of the atmosphere because we assume it is clear, dry, and cool. We must also assume that objects emit energy uniformly (e.g., Lambertian objects) in order to use Planck's Law to model the direct contribution. The camera spectral response is measurable, and an estimate for a similar FLIR sensor was available[1]. Finally, we assume that the camera lens has perfect transmission and no falloff. These last two assumptions are false. However, these and some of the other effects we are assuming to be negligible could be accounted for in the future either by including them in the calculations explicitly, or with a technique such as style transfer[185] or image-to-image translation[345].

Given these assumptions, we model the signal at the sensor using only the direct contribution, given by Planck's Law (Eq. 5.1):

$$L(T, \varepsilon_{avg}, R_\lambda) = \varepsilon_{avg} \int_{\lambda=8\mu m}^{\lambda=14\mu m} R_\lambda \left(\frac{2hc^2}{\lambda^5} \frac{1}{\exp(\frac{hc}{kT\lambda}) - 1} \right) d\lambda \tag{5.1}$$

where L is radiance $[W/m^2/sr]$, T is temperature [K], ε_{avg} is the average emissivity over the bandpass, R_λ is the peak normalized camera spectral response, h is Planck's constant, c is the speed of light, λ is the wavelength $[\mu m]$, and k is the Boltzmann constant. Emissivity, a value ranging be-

	Winter	Summer	
Object	Temp. (K)	Temp. (K)	Avg. ε
Soil	278	288	0.914[301]
Grass	273	293	0.958[301]
Shrub	273	293	0.986[301]
Acacia Tree	273	293	0.952[12]
Human	292	301	0.985[203]
Elephant	290	298	0.96[260]
Zebra	298	307	0.98[199]
Rhinoceros	291	299	0.96
Hippopotamus	290	298	0.96
Crocodile	295	303	0.96
Water	273	293	0.96[203]
Truck	273	293	0.80

Table 5.1: Approximate temperatures and emissivities over night.

tween 0 and 1, relates the radiation of a real object to that of a blackbody, which is a perfect emitter. A blackbody would have an emissivity of 1, and a real object would have an emissivity less than 1. Emissivity is wavelength dependent, but we consider the average over the wavelengths to which the thermal infrared camera is sensitive.

We can calculate this integrated radiance for all objects in our segmentation map from AirSim. For example, given the pixel locations of a human, we can estimate or measure the temperature and emissivity of the human and use Eq. 5.1 to estimate the resulting radiance at the sensor. Table 5.1 contains the temperatures and emissivities that have been estimated for the objects of interest in the African savanna environment that were used for calculations.

5.4.2 Blur and Noise

To this points, we have not considered blur or noise. The point spread function (PSF) is a measure of blur, as it describes the response of an imaging system to a perfect point of radiance. At best, the imaging system will be diffraction-limited, which will lead to some blur around the point of radiance. However, other factors, such as imperfections in the lens or atmospheric effects, can also contribute to the PSF and lead to blur in the image[263]. After light passes through the environment and the lens, it interacts with the detector to create an image. Noise is present in all detectors. Microbolometers are the detectors that are commonly used in uncooled thermal infrared cameras. When a thermal infrared photon strikes the detector, the temperature rises, and the resistance of the detector changes[8]. According to[173], the three main sources of noise in microbolometers are Johnson noise, flicker noise, and thermal noise. The Johnson noise is due to the resistor nature of the microbolometer. The flicker noise is due to flaws in the material surface in semiconductors[262]. The thermal noise is due to the heat exchange with the environment, which is important with uncooled microbolometers, though can be mitigated by changing the gain.[8] mentions that there is also fixed pattern noise (FPN) due to the fact that each microbolometer has a slightly different resistance for the same incoming thermal infrared photons. Although there are in fact other noise sources, such as periodic noise, which can be present in these videos, we focus on Johnson noise, flicker noise, thermal noise, and FPN. Other noise sources could be incorporated in the future.

In order to model these phenomena, again in a simplistic manner, we first utilize a Gaussian distribution for the PSF. This could be replaced with a real model of the PSF for the cameras being used in the field based on images they capture. However, the Gaussian blur kernel used here to loosely approximate a PSF has a standard deviation of 1, which was chosen visually.

Thermal noise and Johnson noise are both characterized by white Gaussian noise ([173,108]). We utilize Gaussian 1/f noise to model the flicker noise[306]. Both are modeled based on[334,172]. Finally, the FPN is modeled as uniform random noise[8]. The same noise distributions were used for all frames of the same video, with the FPN scaled by the first image's standard deviation. All are added to the normalized image, which is then scaled and clipped, to produce the final image.

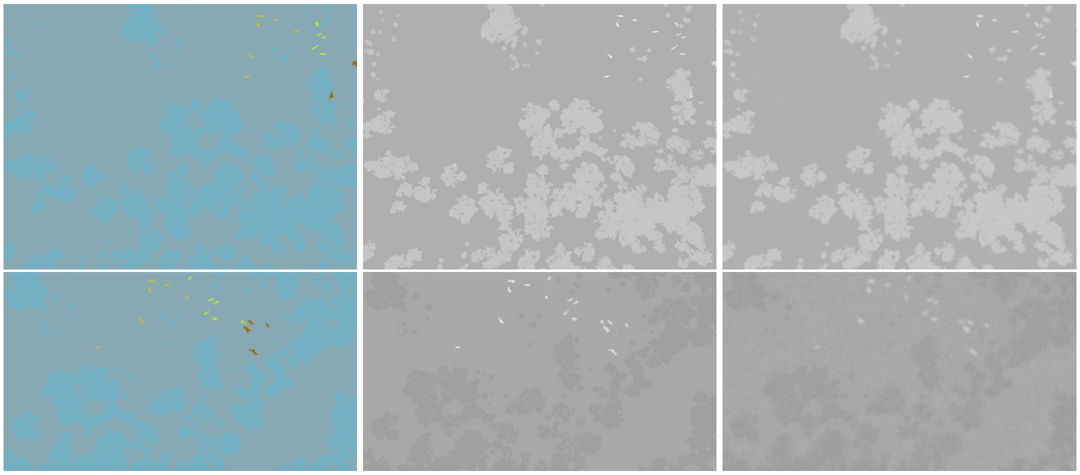


Figure 5.2: Segmentation, thermal infrared image without noise, and final thermal infrared image. Top: summer, bottom: winter. Both rows contain animals.

5.4.3 Process

In order to convert from RGB to infrared, therefore, we now have the following: a segmentation map from the RGB simulation that specifies the objects in each image captured, a thermal infrared digital count associated with all of the objects in the simulation, and a simple model for blur and noise. We therefore assign the thermal infrared digital count to the corresponding object in the segmentation map to get a thermal infrared image. Finally, we add the blur and noise. Fig. 5.2 shows two examples, one each for winter and summer temperatures, where we see the segmentation map, the corresponding thermal infrared image, and the image with blur and noise.

5.5 Utilizing AirSim-W for Human and Animal Detection

5.5.1 Generating Training Data with AirSim-W

In order to generate simulated thermal infrared imagery for use in deep learning algorithms, we follow the workflow depicted in Fig. 5.3. We utilize the Python API and add the option to fly in a

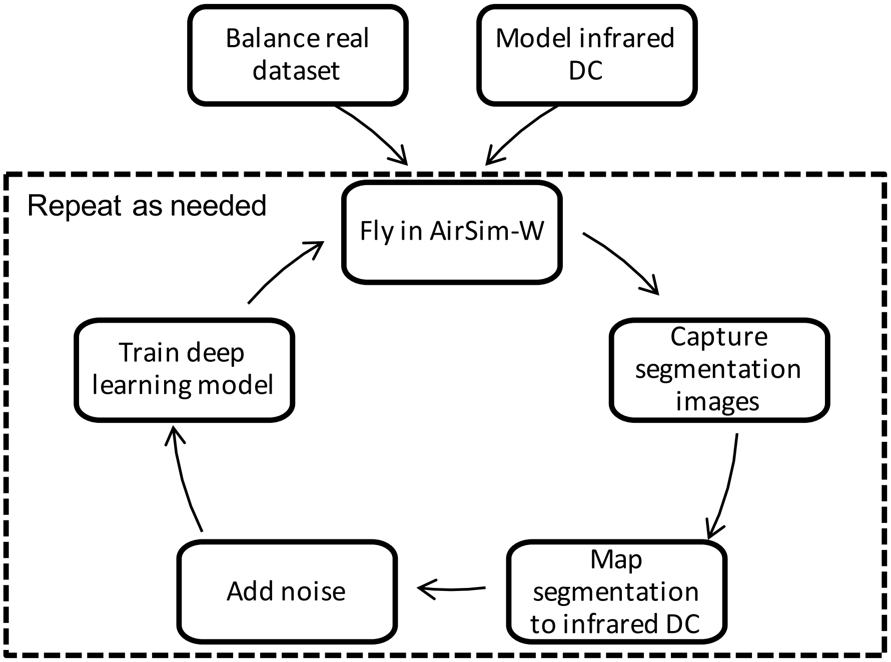

Balance real dataset
Model infrared DC
Repeat as needed
Fly in AirSim-W
Capture segmentation images
Map segmentation to infrared DC
Add noise
Train deep learning model

Figure 5.3: Workflow for generating deep learning training data with AirSim-W, particularly for generating data for human and animal detection in thermal infrared data.

zig-zag pattern, or to return a position for a specific object of interest at each time step. This could then be used to follow the specific object of interest, such as a human, to ensure the object is in the frame at all times. Furthermore, we adjust flight altitude and look angle using Computer Vision Mode, and we adjust the season to determine which digital counts should be used. Once these parameters are set, we fly, either in the zig-zag pattern or following an object of interest, and capture the segmentation image in each time step. Finally, we convert this image into the thermal infrared image for the time step.

For evaluation purposes, this process was used to generate data from 12 flights, 6 summer and 6 winter. Together, this yielded 84,073 individual frames containing objects of interest. Each of the 12 flights consisted of 30 minutes of flying time, totaling 6 hours.

5.5.2 Balancing Data

In the case of real thermal infrared videos of animals and humans captured aboard conservation drones, we have six classes: small animal, medium animal, large animal, small human, medium human, and large human, where the threshold for "small" is an average bounding box area of about 200 pixels and less throughout the video, and "large" is on average greater than 2000 pixels. These are balanced with negative samples automatically in Faster RCNN. However, because we have full videos which must stay consistently in either the test or train sets, and individual frames that may contain multiple objects of interest from different classes, we cannot simply randomly sample full videos for use in training, as was done in[38]. If we do, we may not actually balance all six classes in terms of individual samples. For example, if we have 100 frames with two small humans and one small animal, we must take into account that there will be 200 small humans introduced by including all 100 frames, while there will only be 100 small animals introduced. Furthermore, we would like to be able to detect at all three sizes.

Therefore, we sample the training set through the use of a mixed-integer linear program (MILP). The motivation of using a MILP is that we would like to use as many different train videos as possible, so the balanced dataset will not just sample all consecutive frames from just one or two videos. In other words, by utilizing more unique videos, we provide our algorithm with samples with more variety. We also define "frame types", each can be represented by a 6-dimensional vector, indicating the number of objects from each class. For example, one frame type could be $(1, 0, 0, 1, 0, 0)$ meaning frames of this type have one small animal and one small human. A video can have frames of various frame types. For simplicity in the paper, we denote the frame types as type 1, 2, 3, etc. Therefore, our objective is to use frames from as many different videos as possible, while maintaining balance between the total number of labels in different classes, and bearing in mind that we have many different frame types in videos.

We now formally define this as an MILP. i is the index of the video, j is the index of the frame type, and k is the index of the label type (i.e., which class the object of interest belongs to). Then, c_{ij}^k is the number of type k labels in the type j frame in video i (e.g., if type 2 frame is "empty" frame in video 1, then $c_{12}^k = 0$ for any k, if type 4 frame is "single small human only" frame in video 1, then $c_{14}^1 = 1$ and $c_{14}^k = 0, \forall k \neq 1$). N_{ij} is the number of type j frames in video i. L_l^k and L_u^k are the lower bound and upper bound, respectively, of the desired total number of type k labels. These bounds implement the balance requirement on the total number of labels in different classes.

x_{ij} is a variable representing the number of type j frames in video i that are sampled or selected. u^k and v^k are variables referring to the maximum and minimum number of type k labels that are selected from a single video among all videos except the videos that have no type k label at all and the videos whose type k labels are all selected (i.e., $\sum_j c_{ij}^k x_{ij} = \sum_j c_{ij}^k N_{ij}$). Finally, w_i^k is binary indicator indicating if all type k labels in video i are selected.

$$\min \sum_k (u_k - v_k) \tag{5.2}$$

$$u^k \geq \sum_j c_{ij}^k x_{ij}, \forall i, k \tag{5.3}$$

$$v^k \leq \sum_j c_{ij}^k x_{ij} + M w_i^k, \forall i, k \tag{5.4}$$

$$w_i^k \in \{0, 1\}, \forall i, k \tag{5.5}$$

$$M(1 - w_i^k) \geq \sum_j c_{ij}^k N_{ij} - \sum_j c_{ij}^k x_{ij}, \forall i, k \tag{5.6}$$

$$x_{ij} \leq N_{ij}, \forall i, j \tag{5.7}$$

$$L_l^k \leq \sum_i \sum_j c_{ij}^k x_{ij} \leq L_u^k, \forall k \tag{5.8}$$

$$x_{ij} \in \mathbb{N}, \forall i, j \tag{5.9}$$

(2) is the objective function, which minimizes the difference in the number of labels of each type selected from videos. The objective function implicitly encourages a balanced number of labels being selected from different videos for each label type. In the ideal case, this objective function takes value 0, when each label type, an equal number of labels is selected from each video. (3) and (4) define the variables in the objective function. In (4), we will multiply by M, a large positive number, if all k labels in the video i are selected, as defined in (5) and (6). (7) simply ensures that the number of sampled frames is less than or equal to the number of frames in the video i. (8) ensures that we are within the desired number of samples based on our frame choices, and finally (9) ensures the number of frames sampled is integer.

The introduction of w_i^k is to make sure that when we compute v_k, we exclude the videos whose type k labels are already fully selected. Consider the case where there is a video A that only has 3 large animal labels in total among all frames. Then, we want to include all these labels in our selection, and at the same time, we also want to balance the number of labels of large animals in other videos which have a large number of large animal labels. So, we need to exclude video A when computing v_k.

The current MILP enforces this requirement. Given the objective function in (2), the optimizer will try to make v_k as large as possible. Since (4) is the main restriction for v_k, the optimizer will try to set w_i^k to be 1 whenever possible. (6) ensures that w_i^k can take the value of 1 only if all type k labels in video i are already selected. Note that setting $w_i^k = 0$ is still feasible when the condition is satisfied, but setting $w_i^k = 1$ can achieve at least the same and sometimes better objective value.

Given the optimal solution to this MILP, we randomly select the specified number of frames from each frame type in each video to achieve a balanced dataset that uses as many different videos as possible.

5.6 Evaluation

5.6.1 Qualitative Tests

First, we examine the simulated images qualitatively. In Fig. 5.4, we observe three pairs of real and simulated frames side-by-side. Although noise has been modeled simply, meaning some periodic noise and gain fluctuations are not present, they otherwise look very similar when it comes to relationships between the objects of interest, such as trees and soil. For example, in Fig. 5.4b, the trees are darker than the surrounding ground, as in the simulation. The same is true for Fig. 5.4a. In Fig. 5.4c, humans are approximately the same size in both.

5.6.2 Quantitative Tests

Qualitatively, the simulated images look similar to the real images (other than noise). We now evaluate the simulated data quantitatively by utilizing it in one wildlife conservation task of interest, detecting humans and animals in thermal infrared images. First, we examine the effects of balancing the real dataset based on the MILP we presented earlier. For the new balanced dataset, we have 651 total labeled frames to train one model, out of the original 39,380 frames. We do not choose different models (e.g., different training data) for animals and humans based on performance as we did previously with SPOT. To conduct this first test of balancing data only, we initialize using pre-trained ImageNet weights for Faster RCNN, and fine-tune using the 651 balanced frames. These results are found in the column labeled "None, Regular" in Tables 5.2 and 5.3, as there was no simulated data used in this initial test of balancing data only.

Next, we examine the effects of adding simulated data. We test two types of simulated data: regular simulated data, without blur or noise added, and noisy simulated data, including the blur and noise discussed in Section 5.4.2. To run these tests, we initialize using pre-trained ImageNet weights for Faster RCNN as before. We then fine-tune using the simulated data, and finally fine-tune using real data. For real data, we conduct two tests: (i) fine-tune using the balanced real dataset, and (ii) fine-tune using the SPOT datasets. SPOT is our previous system based solely on real data. Again, the first test (i) is fine-tuning with balanced data after first fine-tuning with simulated data, and the second test (ii) is fine-tuning with the SPOT unbalanced data after first fine-tuning with simulated data. Each fine-tuning process takes 4 hours on an NVIDIA Titan X (Pascal).

The results for fine-tuning using the simulated data only, without noise, can be found in the column labeled "Regular, None", the results for (i) are labeled "Regular, Balanced" and "Noisy, Balanced" based on the type of simulated data, and the results for (ii) are labeled "Regular, SPOT". We also include the previous results from SPOT in the first column. Again, SPOT uses different models for human and animal videos, so results for SA, MA, LA for "None, SPOT" and "Regular, SPOT" are fine-tuned using the SPOT animal model, and the results for MP, LP are fine-tuned using the SPOT human model. Note that the simulated data is primarily balanced by construction because humans and animals are co-located in the simulation environment, and because we believe that balancing the data becomes less important when there is a large amount of it. Also note that in the simulated dataset, we flew only at 200 and 400 ft, which means there were no large animals or large humans, and most objects of interest were actually around the small-medium data threshold of 200 pixels in area.

The test set contains six historical videos containing animals or humans of different sizes. These are the same test videos used to evaluate SPOT in[38]. The combined results for all tests can be found in Table 5.2 and Table 5.3. SA, MA, LA, MP, and LP represent the objects of interest in that particular test video. They are small animals, medium animals, large animals, medium humans, and large humans, respectively. Small humans were excluded because, as with SPOT, none in this particular test video were identified correctly. This is because the human bounding boxes are less than 20 pixels in area.

Table 5.2: Precision results.

Simulation	None	None	Regular	Regular	Regular	Noisy
Video/Real	SPOT	Balanced	None	SPOT	Balanced	Balanced
SA	**0.5729**	0.5232	0.0166	0.4044	0.3536	0.4286
MA	**0.5544**	0.5510	0.0041	0.5066	0.5228	0.5498
LA	**0.5584**	0.3873	0.0318	0.5407	0.4404	0.4592
MP	0.0995	0.1660	0	0.1136	0.1864	**0.2633**
LP	0.3977	0.3571	0.0074	**0.7799**	0.2286	0.0294
Avg.	0.4366	0.3969	0.0120	**0.4690**	0.3464	0.3461

Table 5.3: Recall results.

Simulation	None	None	Regular	Regular	Regular	Noisy
Video/Real	SPOT	Balanced	None	SPOT	Balanced	Balanced
SA	0.0025	0.0026	**0.0044**	0.0027	0.0020	0.0014
MA	0.0131	0.0278	0.0117	**0.0355**	0.0272	0.0254
LA	0.2293	0.2939	0.1297	0.2825	**0.3149**	0.2971
MP	0.0073	**0.1304**	0	0.0111	0.1168	0.0953
LP	0.0188	0.0054	0.0038	**0.0374**	0.0014	0.0004
Avg.	0.0542	0.0920	0.0299	0.0738	**0.0925**	0.0839

5.6.3 Discussion

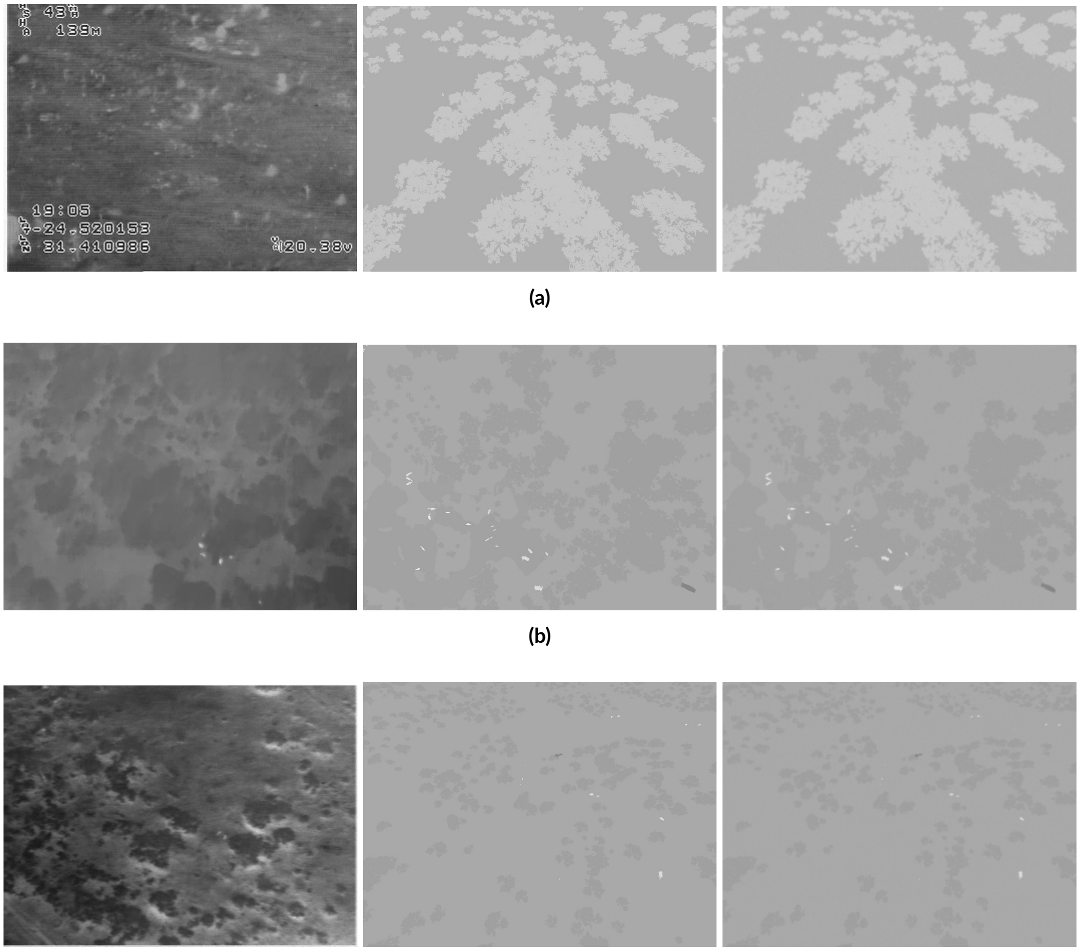
There are several interesting results. First, for recall, using simulated data produces best results for 4 test videos and on average. It is especially interesting that using only simulated data without any real data produces the best recall results for SA. Using simulated data plus SPOT produces the best precision results on average, though SPOT without simulated data does produce the best precision results overall for videos SA, MA, and LA. We believe this could be attributed to several reasons: (i) we selected the model from SPOT that performed best with animals and kept this separate from the human model, (ii) using more real data is better than using more simulated data in general, as the SPOT animal model used about 18,480 real animal frames, or (iii) we lacked large animal examples in simulation. More simulated data could be generated in the future to test this.

The addition of noise only improves over the other datasets in the case of precision for MP. This is interesting, as it implies that perfect images for initial training may actually be beneficial, or that a more sophisticated noise model such as a GAN is necessary. We can further examine this in future work.

It is also interesting to note that balanced data alone performs comparably to SPOT, which used 22,663 frames, while balanced data used only 651 frames. This implies that having a dataset with variety might mean that less data is needed. For example, if we must label real data, we may consider labeling only a few frames per video in the future as opposed to labeling full videos in[39]. We may also consider different distinctions than small, medium, and large, or assign these distinctions per frame instead of on average to further improve balancing. In addition, using simulated data only for fine-tuning while testing on real data does provide nonzero results on most videos, sometimes comparable with real data only. This implies that should labeling a large dataset be too costly, generating large amounts of simulated data may be sufficient to achieve results on real data, and will reduce significant labeling burden. Either of these techniques, or both combined, could allow for less costly, better data collection in the future. Future work could determine the optimal amount of simulated and real data.

5.7 Conclusion

In conclusion, we present AirSim-W, a new simulation environment and data augmentation technique built specifically for wildlife conservation. AirSim-W includes the (i) creation of an African savanna environment in Unreal Engine, (ii) thermal infrared modeling, (iii) new methods to fly the conservation drones throughout the scene for training data collection, and (iv) demonstrated detection improvement using simulated data generated by AirSim-W. Labeling real data costs over $8000, while the creation of the simulated environment, which can generate unlimited amounts of data, costs closer to $5000. The cost of the simulated data could be lowered in the future when expanding to other animals and environments by developing transfer learning techniques, possibly by using the existing Africa environment (https://github.com/Microsoft/AirSim/releases), and/or by finding existing environments and animals. Also, labeling real data took approximately 800 hours total, whereas creating the environment and generating simulated data took approximately 200 hours. With these contributions, AirSim-W will be a cost efficient, useful tool for wildlife conservation research, especially for the problems of human and animal detection in conservation drone videos, patrol planning for conservation drones and foot patrols, and camera trap placement.



(a)

(b)

(c)

Figure 5.4: Qualitative comparison of real frames (left), basic thermal infrared simulated frames (middle), and noisy simulated frames (right). 5.4a: summer, 5.4b: winter, 5.4c: winter. The images in the first and third rows contain humans, and the images in the second row contain animals. The simulated images in the third row also contain animals.

6

BIRDSAI: A Dataset for Detection and Tracking in Aerial Thermal Infrared Videos

6.1 Introduction

Recent advances in deep learning have led to immense progress in vision applications like object recognition, detection, and tracking. One of the key factors driving this progress is the availability

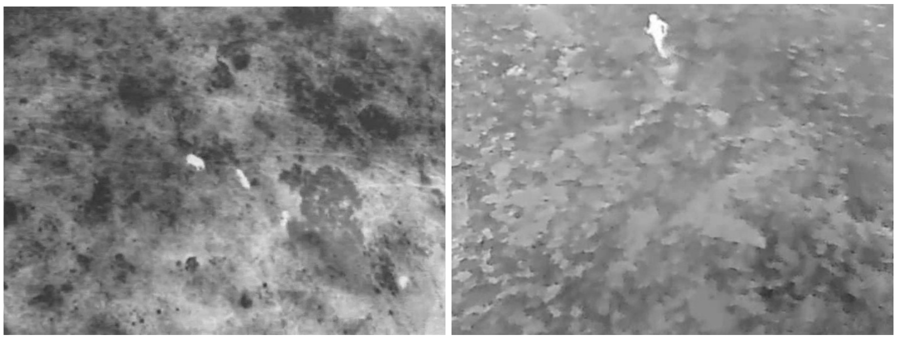


Figure 6.1: Example images from BIRDSAI: elephants and a human, respectively, from an aerial perspective.

of large-scale datasets capturing real-world conditions along with careful annotations for training and comprehensively evaluating machine learning models. The collection and release of many of these datasets is often inspired by specific applications of interest, e.g., perception for autonomous driving using object detection, tracking, and semantic segmentation, person re-identification for surveillance camera networks, and facial recognition for biometrics and security applications. While the majority of the publicly available datasets cater to techniques developed for the visible spectrum[84,90,188,162,150,75,103,321,87], there has been an increasing interest in applications from the near-infrared (NIR) and thermal infrared (TIR) spectral ranges[320,27,180,131,176], as these sensors become more affordable.

Concurrently, with advances in aerial image acquisition technology, datasets specifically targeting object detection and tracking in aerial images have been made publicly available[177,321,87]. In[321], the images have been acquired from various remote sensing sources (e.g., satellites), and capture varying degrees of orientation, scales, and object density. On the other hand, aerial images from drones[177,87] are often motivated by applications like monitoring, yet these images are restricted to the visible spectrum, thereby limiting their usage to well-lit conditions. Besides, most existing public datasets, aerial and terrestrial alike, address applications relevant to relatively densely populated settings.

Contrarily, our work is motivated by recent concerns about depleting biodiversity and loss of forest cover which are exacerbated by illegal activities such as poaching for wildlife trade, hunting,

Dataset (Year)	Platform	#Frames	Tasks	Spectrum	(R)eal
UTB[177] (2017)	A	15K	S	V	R
UAV123[209] (2016)	A	113K	S	V	R,S
UAVDT[87] (2018)	A	80K	D,S,M	V	R
TIV[320] (2014)	G,A[a]	64K	D,S,M	T	R
LTIR[27] (2015)	G,A	12K	D,S,M	T	R
PTB-TIR[180] (2018)	G,A	30K	D,S,M	T	R
ASL-TID[242] (2014)	A[a]	5K	D,S,M	T	R
[329] (2015)	G,A[b]	84[c]	RE	T,V	R
[190][d] (2016)	A	9K	D,S,M	T	R
BIRDSAI (proposed)	A	62K + 100K	D,S,M	T	R,S

Table 6.1: Comparison summary of recent aerial video datasets for detection and tracking. Platform could be either (A)erial or (G)round-based; #Frames is the total number of annotated frames in the dataset, with our dataset reporting 62K (1K=1000) real frames and about 100K synthetic frames; Tasks for which annotations are present (D)etection, (S)ingle-object, (M)ulti-object tracking, and (RE)gistration; Spectrum of cameras: (V)isible or (T)hermal-IR; Data acquisition (R)eal or (S)ynthesized in a simulator. Comparisons are discussed in Sec. 6.2. [a]Fixed aerial perspective; [b] Aerial images do not contain humans or animals; [c]84 Pairs; [d] Not publicly available, contains primarily images of roads, and has portions of images used for tracking.

and logging. Efforts to mitigate these activities through patrolling of protected areas, especially at night, is very challenging and puts forest rangers at risk due to poor visibility, difficult terrain, and increased predator and poacher activity[226]. These conservation efforts are increasingly being augmented by conservation drones[226,153,146,6], with TIR cameras as the preferred sensing modality for night-time monitoring over natural landscapes where the ambient light is minimal and the conservation drone's altitude, capacity, and need for stealth preclude the use of active light sensors. However, manual monitoring of aerial TIR videos to detect and track humans in real time is an extremely challenging and tedious task, especially when the goal is to interdict an illegal activity.

In this paper, we introduce Benchmarking IR Dataset for Surveillance with Aerial Intelligence (BIRDSAI, pronounced "bird's-eye"), a large, challenging aerial TIR video dataset for benchmarking of algorithms for automatic detection and tracking of humans and animals. To our knowledge, this is the first large-scale aerial TIR dataset, with multiple unique features. It has 48 real aerial TIR videos of varying lengths, carefully annotated with objects like animals and humans and their trajectories. These were collected by a conservation organization, Air Shepherd, during their regular surveillance efforts flying a fixed-wing conservation drone over national parks in Southern Africa. Finally, we augment it with 124 synthetic aerial TIR videos generated from AirSim-W[37], an Unreal Engine-based simulation platform. Two example images from real videos are shown in Fig. 6.1 depicting a herd of elephants and a human. Realistic and challenging benchmarking datasets have had tremendous impact on the progress of a research area. Synthetic datasets like[252,258,98] along with real ones like[76,103] have accelerated the progress in unsupervised domain adaptation techniques[171,297]. Similarly, the Caltech-UCSD Bird (CUB-200) dataset[311,307] has helped advance an important area of fine-grained visual recognition[344]. With more wildlife monitoring datasets[281,25,96,317] becoming publicly available, we may expect rapid progress in areas like species detection, counting, and visual animal biometrics[77,96,63,168]. Inspired by these instances, we anticipate the proposed dataset will promote advances in both (i) algorithm development for the general problems of object detection, single and multi-object tracking in aerial videos, and their domain adaptive counterparts, and (ii) the important application area of aerial surveillance for conservation.

The rest of the paper is organized as follows. First, we introduce the gap in existing datasets that we aim to fill with the proposed one (Sec. 6.2). We then discuss the attributes of the dataset in detail (Sec. 6.3), such as the means of acquiring the data, strategies adopted for annotation, and the train/test splits. We next analyze the content of the resulting dataset (Sec. 6.4), and evaluate the performance of well-known techniques for the tasks of object detection, single and multi-object tracking, and domain adaptation (Sec. 6.5) before finally concluding the paper (Sec. 6.6).

6.2 Motivation

With poaching becoming widespread around the world[300], aerial surveillance with conservation drones is becoming a mainstream application[226,153,146]. In order to apply deep learning-based de-

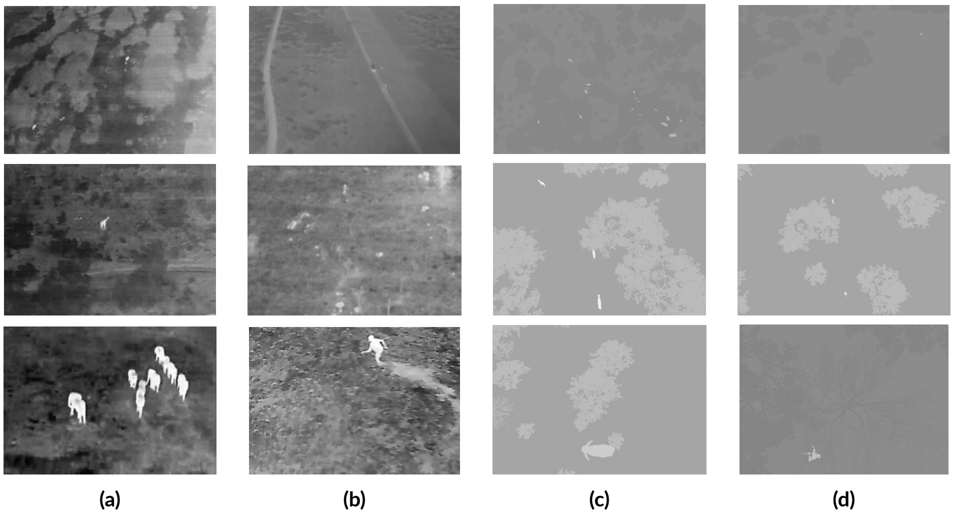


Figure 6.2: Sample images from the real and synthetic datasets. From top to bottom: small, medium, and large objects. (a) & (b) Real images of animals and humans, respectively; (c) & (d) Synthetic images of animals and humans, respectively. Mixture of summer and winter synthetic data (winter has dark trees compared to ground).

tection and tracking techniques to these applications (especially at night) and evaluate performance, there is a need for a realistic, large, annotated dataset that adequately captures the challenges faced in the field. Recently, several large datasets for aerial image analytics have been publicly released, many of which were captured using drones. However, all of these are data in the visible spectrum. In the rest of this section, we discuss some of the most closely related public datasets and highlight the unique aspects of the presented dataset. A summary of comparisons with existing datasets is provided in Table 6.1.

Existing Drone Datasets: The recently introduced UAVDT[87] contains nearly 80,000 frames with over 0.8 million bounding boxes. The dataset is comprised of videos collected over urban areas with object categories of cars, trucks and buses. The DTB dataset[177] was introduced for benchmarking drone-based single object tracking with the goal of jointly evaluating the motion model and tracking performance. Mueller et al. introduced the UAV123 dataset[209], which contains 123 HD video sequences with about 113,000 annotated frames captured by a low-altitude drone. Eight of these videos were rendered using an Unreal Engine environment. All of these datasets use visible spectrum cameras mounted on multirotor drones, which typically have lower speeds and better image stabilization as compared to fixed-wing drones[44]. In a poaching prevention application, deploying a multirotor drone for surveillance is more difficult due to stealth and coverage requirements.

Existing TIR Datasets: The BU-TIV dataset[320] is part of the OTCBVS dataset collection* and contains 16 video sequences with over 60,000 annotated frames for tasks like detection, counting and tracking. The LTIR[27] dataset was used for the VOT-TIR 2016 challenge and contains 20 video sequences of length 563 frames on average. The PDT-ATV dataset[242] was introduced for benchmarking tracking of pedestrians in aerial TIR videos. All eight sequences are captured using a handheld TIR camera at a height and angle to simulate a drone, but because it is handheld, it is a fixed aerial perspective. Recently, the PTB-TIR dataset[180] was also introduced for benchmarking TIR pedestrian tracking. It is comprised of 60 sequences with over 30,000 annotated frames. In all cases, the challenge of analyzing TIR footage *from a conservation drone* has not been addressed yet.

Synthetic Datasets and Domain Adaptation: Training deep models demands large datasets with accurate annotations, which are expensive and tedious to obtain. Consequently, recent years have seen an increasing use of synthetic images rendered using state-of-the-art graphics engines, where generating accurate ground-truth information is trivial. Several such simulators and datasets have been made public[98,252,258,210], and reportedly improve real-world performance of deep learning models when pretrained with synthetic data. The increased use of synthetic datasets in vision applications has further propelled research in domain adaptation with works like[64,171,297] being a few among the many† leveraging synthetic data. Some recent work has also shown that domain transformation and adaptation techniques can aid in improved detection performance in TIR images when using deep CNNs pretrained on visible spectrum images[66,28]. In contrast, BIRDSAI uses both real

*http://vcipl-okstate.org/pbvs/bench/

†More comprehensive list of advances in domain adaptation: https://github.com/zhaoxin94/awsome-domain-adaptation

and synthetic aerial, TIR videos.

BIRDSAI: The 48 real TIR video sequences included in BIRDSAI were randomly selected from a database of conservation drone videos collected by Air Shepherd for conservation, and contain 1300 frames on average. These videos accurately reflect the challenges in the field, e.g., motion blur, large camera motions (both rotations and translations), compression artifacts due to bandwidth constraints, background clutter, and high altitude flight (60-120m) resulting in smaller objects to detect and track. The 124 synthetic videos with 800 frames on average, on the other hand, were generated using the AirSim-W[37] platform with the publicly available models of the African savanna, animal species, and conservation drone-mounted cameras. This dataset uniquely brings together the three categories discussed above.

6.3 Dataset Description

6.3.1 Real Data

Data Acquisition

Data were collected throughout protected areas in the countries of South Africa, Malawi, and Zimbabwe using a battery-powered fixed-wing conservation drone. Specific locations are withheld for security. All flights took place at night, with individual flights lasting for about 1.5 - 2 hours. Various environmental factors such as wind resistance determined exact flying time. Throughout the night, there were typically 3 to 4 flights, the altitude ranged from approximately 60 to 120m, and flight speed ranged from 12 to 16 m/s depending on conditions such as wind. Temperature ranged from less than 0 to 4 in winter at night, though typically closer to 4. There was often a shift of approximately 5 throughout the course of the night in the winter. For reference, daytime temperatures were typically approximately 15 to 16. During summer, the temperature ranged from 18 to 20 at night, and 38 to 40 during the day. When flying just after sunset, the ground temperature was warm and could make it more difficult to spot objects of interest due to the lack of contrast. However, by about 10:30-11PM, there was typically sufficient contrast for easier visibility. Fog was present in some rare cases, which could cause "whiteouts" in images.

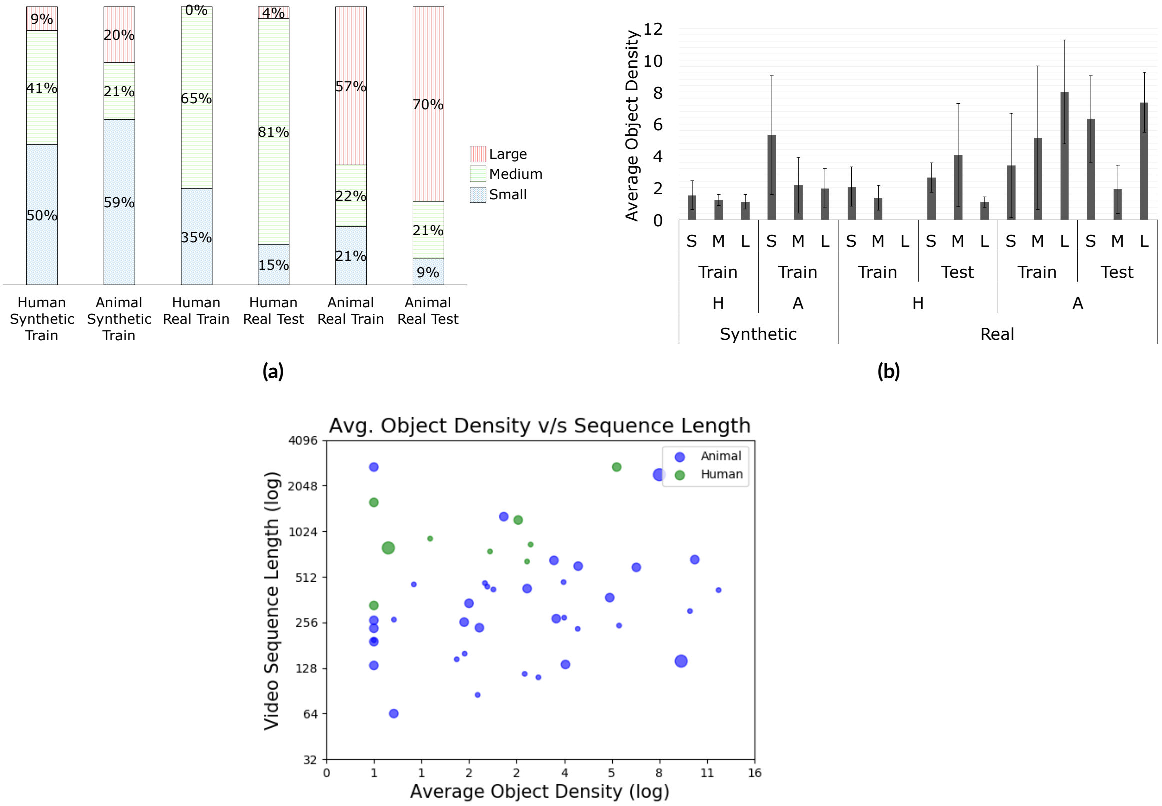
The FLIR Vue Pro 640 was the primary sensor utilized. However, the Tamarisk 640 was also used in some videos in the dataset. Although the typical resolution of images is 640x480 as a result, some images may be sized differently due to the removal of text embedded in to the videos describing specific locations and other flight parameters, which are also withheld for security purposes. These cameras produce 8-bit images and use Automatic Gain Control (AGC), as in[67]. This leads to more reliable contrast that facilitates better detection and tracking accuracy during flight. The cameras cost approximately $2000-$4000 depending on the lenses and other attributes. They have 19mm focal length and collect imagery at a rate of 30Hz. Images were streamed to a base station during flight, where they were stored as raw videos. All videos were converted to mp4 videos for processing and JPEG images. Because the videos were recorded from real-world missions, they lack some metadata, such as speed, altitude, and temperature. While this auxiliary information could be useful, automatic vision algorithms should still be designed to work in their absence. From a usability perspective, this added robustness is crucial for building practical vision systems that are less sensitive to specific conservation drone or camera settings.

Annotation

We used VIOLA[39] to label detection bounding boxes in the thermal infrared imagery, and followed the process described in VIOLA. To briefly summarize this labeling process in VIOLA, after labels were made by one person, two other people reviewed the labels, making corrections as needed. General rules that were followed during the labeling process are as follows. If individuals were completely indistinguishable (e.g., multiple humans or animals were close together and could not be distinguished at all in thermal imagery), they were not labeled. Instead, occlusions are recorded when possible to determine manually from context. This includes cases where animals or humans become indistinguishable for a few frames and again become distinguishable after they or the camera move. If there were artifacts in the image (see Sec. 6.4), objects were tagged as containing noise. Some extremely small amounts of these artifacts may have been allowed without being tagged as noisy. We provide examples of how we included occlusion and noise in the Appendix. Finally, if an object was mostly out of the camera's field of view (i.e., more than about 50% of the object was not present in the frame), it was not labeled. After this process, all labels were finally confirmed and checked for quality for use in this dataset by the authors, one of whom is from Air Shepherd and collected the videos, for a total of 4 checks on each initial label.

We additionally labeled individual species when distinguishable, typically in videos with larger animals present. The real videos contain giraffes, lions, elephants, and a dog, which account for about 100K of the 120K individual animal bounding boxes (the remaining 20K animals are marked as unknown species). There are about 34K human bounding boxes. These labels created using VIOLA were then labeled separately for tracking. We built a tool using Tkinter[‡] to assign object IDs to each bounding box label. To reduce annotation effort before any human annotation was done, the tool checked for overlap between frames using an Intersection over Union (IoU) threshold. If the IoU exceeded the threshold, the object in the following frame was given the same object ID. Once this automatic processing was complete, we used the tool to manually navigate through the video frames and identify and correct any errors in the assigned object IDs, e.g., objects merging or splitting. In the case of objects merging together, object IDs are maintained whenever it is possible to distinguish them again after the merge. However, if they enter a large group, it may become impossible to distinguish which animal is which due to the nature of thermal imagery. In these cases only, they are assigned a new object ID. If objects leave the frame, they will similarly retain the same object ID if possible.

[‡] https://docs.python.org/3/library/tkinter.html



(a)

(b)

(c)

Figure 6.3: Statistics of real and synthetic data. (a) 100% stacked bar charts of distribution of small, medium, and large animals/humans across real and synthetic data and train/test sets. Real train contains 32 videos, real test contains 16 videos, and simulated train contains 124 videos. (b) Bar plot (with standard deviation error bars) of the number of animals and humans for train/test sets over large, medium, and small objects, again across real and synthetic data and train/test sets. (c) Scatter plot showing different video sequences plotted using their constituent average object density (#objects/frame) and sequence length (duration for which the objects were visible in the video). The color indicates the constituent object type (human/animal) and the size of the circles indicate small, medium, or large. For better visual clarity, both the axes are plotted using the log scale.

6.3.2 Synthetic Data

To generate synthetic data with AirSim-W, we utilized the African savanna environment introduced in[37]. In brief, the environment is not based on a particular area of interest, but rather represents the variety of environments found in Southern Africa, such as wide-open plains to dense forest, flatland to mountainous terrain, roads, and water. Grass in the plains is not a mesh in the environment, so in the segmentation provided by AirSim-W, grass and soil are indistinguishable. This does, however, increase efficiency while running the simulation. The AirSim-W platform has a TIR model that was introduced in[37]. We used this TIR model to generate images of the objects in the scene as the conservation drone flew through the environment and captured images of size 640x480. Specifically, AirSim-W's Computer Vision Mode was used, and the conservation drone was placed by following certain objects in the environment. For example, to generate human images, the conservation drone tracked the human. Because the objects move in groups, and multiple altitudes, offsets, and camera angles were used, multiple objects or few objects may have been captured. Ground truth object IDs and species (lions, elephants, crocodiles, hippos, zebras, and rhinos) labels were also recorded for a total of about 220K individual animal bounding box labels and 50K human labels.

6.3.3 Train and Test Sets

In order to create the train and test sets for the real data, our goal was to create similar distributions in both while ensuring complete videos stayed entirely in either the train or test set. Entire videos remained in one or the other because consecutive frames could be extremely similar. We manually assigned videos to the train or test set based on the number of objects in the video, and based on characteristics of the videos, like contrast, to try to ensure an approximately even distribution in the train and test sets. Because entire videos needed to stay together, it was not possible to maintain exact ratios. In fact, there was only one video that contained large humans, so it was placed in the test set only. These train and test sets are shown in Fig. 6.3.

Regarding the synthetic dataset, the entirety of the dataset was used for training. Although we attempted to ensure the approximate ratio of humans and animals was somewhat similar to the real training dataset, we prioritized adding large human examples and more small human and animal examples (see Section 6.4.1 for more description of scale) while generating the synthetic dataset, as these were less frequently seen in the real data. Different statistics over the entire dataset, including distribution of object scales and densities across the train/test splits, are shown in Fig. 6.3 (a) and (b), respectively. In Fig. 6.3 (c), a scatter plot of tracking video sequences is shown with respect to the sequence length and average object density.

6.4 Dataset Properties

The real and synthetic data contain significant variations in content and artifacts, including scale and contrast. The real data also contain more background clutter and noise.

6.4.1 Content

Environments. There are several types of environments that are captured in the dataset, including land areas with varying levels of vegetation and water bodies, such as watering holes and rivers. An example of water with a boat floating upon it is shown in Fig. 6.4 (b) (where the bright, top right portion of the image is water). We denote the presence of water for individual videos in the dataset.

Scale and Density of Objects. There are multiple scales of objects in the dataset. We coarsely categorize them into small, medium, and large based on each object's annotated bounding box area and dataset statistics. These distinctions are assigned to full videos based on the average bounding box size throughout the video[§]. There is also a wide range of densities in objects throughout the videos. The average number of objects per frame (density) for small, medium, and large videos is described in Fig. 6.3. There is an example of a video with high animal density in Fig. 6.4 (a).

[§] Small videos were those whose average bounding box area was $<$ 200 pixels, the median real area, and large videos were $>$ 2000 pixels.

6.4.2 Artifacts

Contrast. Contrast refers to the variation in digital counts in an image. TIR images rely on AGC, so contrast can vary significantly across the dataset. As an example, some images have nearly black backgrounds with white objects of interest (more contrast, e.g., Fig. 6.4 (b)), while others have gray backgrounds (less contrast).

Background Clutter. There can be many objects in the background in some images, particularly in images with vegetation. Vegetation can often have a similar temperature to objects of interest, leading to images like Fig. 6.4 (c). We also see thermal reflections off the ground, typically near trees, e.g., in Fig. 6.4 (d). Both make it challenging to distinguish between objects of interest and background clutter.

Noise and Camera Motion. While there are many sources of noise in TIR cameras that use uncooled microbolometer arrays as the sensor[37,263], the most common type in BIRDSAI is what we call ghosting, as shown in Fig. 6.4 (e). There are also slightly more mild versions of it, which look like horizontal "bands" in some cases. Additionally, the conservation drone's motion, or even the camera motion when there is pan or tilt, can sometimes lead to frames with motion blur. An example of this is shown in Fig. 6.4 (f). These were labeled as containing noise when possible (see Sec. 6.3).

6.5 Evaluation

The goal of BIRDSAI is to advance image-based object detection, domain adaptive detection, and single and multi-object tracking (SOT and MOT, respectively). To evaluate state-of-the-art object detection methods and domain adaptation on BIRDSAI, we perform *framewise* detection of animals and humans. We evaluate tracking by using the videos, both full sequences and subsequences. We provide benchmarking results for these tasks with existing algorithms, leaving the method details

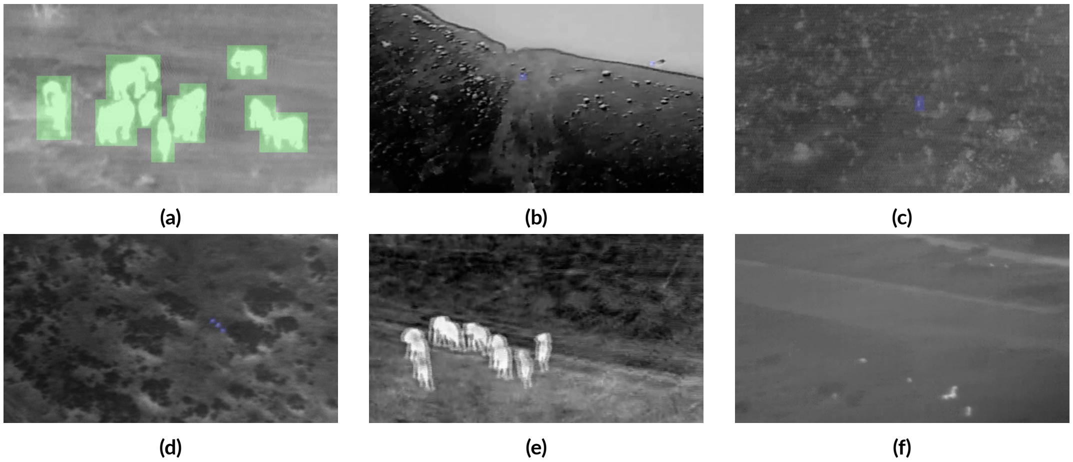

(a) (b) (c) (d) (e) (f)

Figure 6.4: Data challenges. (a) density (b) high contrast (c) clutter (vegetation) (d) clutter (reflections) (e) ghosting (f) motion blur. Ground truth labels not shown in (e) and (f) for better visualization of effects of noise. Animals in (a), (e), (f), humans in (b), (c), (d).

Scale	**FR-CE**	**FR-WCE**	**YOLOv2**	**SSD**
SA	0.216	0.228	0.144	0.182
MA	0.459	0.468	0.383	0.392
LA	0.879	0.896	0.679	0.850
Animals	**0.659**	**0.671**	**0.489**	**0.587**
SH	0.214	0.206	0.108	0.219
MH	0.174	0.179	0.146	0.229
LH	0.154	0.094	0.083	0.147
Humans	**0.181**	**0.155**	**0.104**	**0.183**
Overall	**0.430**	**0.438**	**0.304**	**0.388**

Table 6.2: Detection performance baseline using the mAP metric for different scales ((S)mall, (M)edium, (L)arge) of objects ((A)nimals, (H)umans) in the dataset.

Configuration	**DA-FR-CE**	**DA-FR-WCE**	**FR-CE**	**FR-WCE**	**YOLOv2**	**SSD**
Real → Real	–	–	0.430	**0.438**	0.304	0.388
Syn → Real	0.443	**0.459**	0.309	0.313	0.152	0.294

Table 6.3: Detection performance baselines using the mAP metric after domain adaptation.

to the papers while listing the hyperparameters used for the experiments here. We include further experiments and analyses in the Appendix including cross-dataset evaluation.

6.5.1 Framewise Detection

We specifically test with the following popular object detection methods: Faster-RCNN[250], YOLOv2[248], SSD[182], and Domain Adaptive Faster-RCNN[64], all of which have shown strong results in the visible as well as TIR. Results for detection and unsupervised domain adaptative detection are provided in Tables 6.2 and 6.3, respectively.

Faster-RCNN[331,250]. The experiment was performed using VGG16 as the backbone network initialized with ImageNet pretrained weights. Evaluation results of two loss functions were compared and tabulated in Table 6.2, namely, Cross Entropy (CE) and Weighted Cross Entropy (WCE), to account for the imbalance in the two classes (i.e., humans are more rare). The weights for each of the classes are computed as follows for the WCE loss.

$$W_\ell = \left(\frac{\sum_{i=1}^{k_a} w_i h_i + \sum_{i=1}^{k_h} w_i h_i}{\sum_{i=1}^{k_\ell} w_i h_i} \right)^{0.5} \tag{6.1}$$

where k_a and k_h are the number of animals and humans in the frame, respectively, $w_i h_i$ is the area of the corresponding bounding box, and $\ell \in \{a, h\}$ (animal or human).

These experiments were performed with a batch size of 1, an SGD optimizer, and a starting learning rate of 1e-2. The learning rate was depreciated after a learning step which was set to two epochs for the loss to converge. The overall fine-tuning was performed for a total of 7 epochs.

YOLOv2[248]. This model used pretrained Darknet19 weights. A batch size of 1 was taken with a starting learning rate set to 1e-3, depreciating it by a factor of 10 after every second epoch. The model converges after 12 epochs.

SSD[182]. This model used pretrained VGG16 weights. The hyperparameters of the training include a batch size of 8, initial learning rate of 1e-5, without depreciating the learning rate throughout. The training converges after a total of 12 epochs. An SGD optimizer was used with a weight decay of 1e-4 and an update gamma of 0.1.

Domain Adaptive Faster RCNN[64]. This framework was trained with the base architecture as VGG16, pretrained with ImageNet. The corresponding overall mAPs are tabulated in Table 6.3. Real → Real indicates that the train set is comprised only of labeled real data and the model was tested on real data, which is equivalent to results in Table 6.2. Synthetic → Real implies that the train set is comprised of labeled synthetic data and unlabeled real data, and the testing was performed on the test set (real data).

Results: It is not surprising to note that the best overall performance is achieved using Faster-RCNN with WCE, where the weights explicitly account for the data imbalance. For human objects alone, however, SSD and Faster-RCNN (without weighting) perform comparably, while outperforming the other methods. YOLOv2 performs worst overall, possibly due to the small size of objects. In all cases, there is room for improvement, especially for small animals and humans.

The overall results of Table 6.2 are equivalent to the Real → Real row of Table 6.3. In the Synthetic → Real row, simply training on synthetic data and testing on real data actually decreased performance for those algorithms lacking domain adaptation. This is not surprising either, given that there is a visible domain shift between the synthetic and real data subsets of BIRDSAI. It is, however, encouraging to see that in the unsupervised domain adaptation setting of[64], there is a noticeable increase in the mAP values. This observation suggests that further research in unsupervised domain adaptation could immensely benefit object detection in aerial TIR videos, given the extremely challenging task of annotating aerial TIR videos.

6.5.2 Tracking

We test both single and multi-object tracking on BIRDSAI, and we report results for all objects regardless of class. In both the tracking settings, we use the same train/test splits as used in object detection. For single-object tracking, video sequences were further split into *perfect subsequences* such that each subsequence had a single target object throughout, with a minimum length of 50 frames. Once there was any interruption in the subsequence, whether due to noise, occlusion, or the object exiting the frame, the subsequence ended. This resulted in a total of 552 subsequences. The train/test splits of SOT subsequences were consistent with that of the videos, i.e., all subsequences from test videos were included in the test set, and similarly for the training set. This means that all subsequences from a given video appeared either in the training set or in the test set, which yielded a train set with 386 and a test set with 166 subsequences. For testing of *full sequences*, we used the test videos to generate 99 sequences of length at least 50 frames, with each sequence starting at the first appearance of an object in the video and ending at its last appearance.

For single-object tracking, we use the Siamese RPN[175], ECO[79] and AD-Net[339] algorithms as benchmarks, and we also use the MCFTS[181] algorithm, which was developed specifically for the related VOT-TIR dataset. These algorithms were then evaluated on the test set using the usual metrics of success rate and precision[319,177]. We evaluated pretrained models of ECO and MCFTS, and retrained Siamese RPN and AD-Net on BIRDSAI. We followed the commonly used one-pass evaluation (OPE) process for single-object tracking[319], which required training of models like Siamese RPN and AD-Net to be done on the perfect subsequences, where every frame had ground truth annotations. During testing, we performed the benchmarking on the perfect subsequences and full sequences. As is typical in OPE, all of the trackers were initialized using ground truth bounding boxes in the respective first frames.

For multi-object tracking we only report the IoU Tracker[34] with default thresholds, and object

Method	Perfect Subsequences		Full Sequence	
	Precision	**AUC**	**Precision**	**AUC**
ECO	**0.8103**	**0.5430**	**0.4842**	**0.2972**
AD-Net	0.8029	0.5331	0.4545	0.2546
MCFTS	0.7194	0.4946	0.3401	0.1886
Siamese RPN	0.0073	0.0093	0.0041	0.0048

Table 6.4: Single Object Tracking Evaluation. Precision is at 20 pixels. "Perfect subsequences" excludes noisy/occluded frames, while "Full sequence" includes them.

Method	Obj Size	Ground Truth Det		F-RCNN Det	
		MOTA	**MOTP**	**MOTA**	**MOTP**
IoU Tracker	S	61.6	**100.0**	-102.4	62.7
	M	**91.3**	98.9	-34.4	66.9
	L	80.6	**100.0**	**13.6**	**68.9**

Table 6.5: Multiple Object Tracking Evaluation. S, M, L is for small, medium, large.

detections provided using (i) ground truth bounding boxes and (ii) Faster-RCNN detection. We use Faster-RCNN for MOT benchmarking due to its superior detection results. We also include other MOT results in the Appendix. The algorithms are evaluated using the MOTA and MOTP evaluation metrics[253], where higher is better. MOTA and MOTP are in the range of [-∞, 100 (%)], and [0, 100(%)] respectively. Although they are percentages above 0, negative values for MOTA imply that the errors (false positives, misses, and mismatches) are more than the ground truth objects to be tracked.

Results: See Table 6.4 for SOT and Table 6.5 for MOT benchmarking. For SOT, Siamese RPN, which relies on one-shot detection, fails to perform reasonably. Performance is promising with the other methods. This seems to be related to the length and cleanliness of the track, as evidenced by the improved performance in subsequences compared to full sequences. However, the real world will require handling videos with imperfect tracks and noise, small objects, and detection initialization, which leaves room for innovation. For MOT, IoU Tracker performs very well for ground truth bounding boxes, while it performs worse when using Faster-RCNN detections in both of the MOTA and MOTP metrics.

6.6 Conclusion

We presented BIRDSAI, a challenging dataset containing aerial, TIR images of protected areas for object detection, domain adaptation, and tracking of humans and animals. In our benchmarking experiments, we noted that state-of-the-art object detectors work well for large animals, however, for humans and small and medium animals, the performance drops substantially. Similarly, while IoU Tracker-based multi-object tracking works well when ground truth detections are provided, the performance drops drastically when a detector's output is used. These experimental results indicate the challenging nature of the real sequences in the BIRDSAI dataset. Fortunately, we saw that baseline domain adaptive detection shows promising improvements by leveraging the synthetic dataset. This observation is crucial, as the annotation effort for noisy TIR videos is enormous, and improved unsupervised domain adaptation techniques can prove to be very useful for achieving competitive detection performance. We hope this dataset will help propel research in this important area. Finally, in addition to facilitating interesting research, this dataset will also contribute to wildlife conservation. Successful algorithms could be used to help prevent wildlife poaching in protected areas and count or track wildlife.

7

Micronutrient Deficiency Prediction via Publicly Available Satellite Data

7.1 Introduction

Micronutrient deficiencies, or the lack of vitamins and minerals required by the body for healthy functioning and development[202], are a widespread (estimated to impact more than 2 billion people worldwide, including 340 million children[152]) public health concern that is unfortunately difficult to measure. These micronutrient deficiencies, hereafter referred to as MND, further drive the global burden of disease but remain difficult to diagnose since the effects often become visible only when the deficiency is severe[304]. From a public health perspective seeking to reduce MND prevalence throughout a population, it is important to identify regions at risk of MND. However, due to the difficulty of diagnosing MND, regions with MND are unclear to public health organizations until direct measurements are made, such as blood draws to measure biomarkers and/or surveys/questionnaires. Unfortunately, these blood draws and surveys are costly and time-consuming, and furthermore, quantifying micronutrient levels in a blood sample requires limited, specialized laboratory equipment, leading to infrequent data collection.

Due to the difficulty in both types of data collection, we seek a new data source that may be more scalable, such as satellite data (i.e., data products derived from raw satellite imagery). This may at first seem unrelated, as MND status is unique to an individual, pertaining to an individuals' nutrition, disease status, and other characteristics which cannot be viewed by satellite. Indeed, prior work applying artificial intelligence (AI) techniques to satellite data, e.g., in estimating crop type[97], often search for features directly observable by satellite. Predicting an indirect feature such as MND prevalence brings additional technical challenges, including choosing relevant satellite data, linking a limited amount of ground truth data from individuals to satellite data to train machine learning models, and supporting interpretability for public health experts.

Contributions: Through our novel system, we establish that satellite data can be used to predict MND at a regional level despite these challenges. In fact, our system is the first to predict MND from a regional level, as measured directly from real-world, ground truth biomarkers, using satellite data. This involves i) aggregating individuals' MND states from biomarker data over geographic regions to align with satellite data, ii) using segmentation to generate custom features of importance, specifically market locations in this case, iii) providing scalablity with automatic feature selection, which performs comparably to expert feature selection, and iv) two prediction paradigms to handle the challenges that arise from limited ground truth data: logistic regression, which also naturally handles the pressing need for interpretability of predictions in the field, and multi-layer perceptron with domain adaptation. Not only does this system achieve good accuracy, but this also results in improved performance compared to the baseline of survey-based predictions. We believe this MND detection system could be broadly applied to other countries where satellite data are available, potentially leading to more information for public health interventions and high societal impact.

7.2 Background and Related Work

AI for Social Impact and Satellite Data: Existing applications of AI related to nutrition include food security, agriculture, food rescues, and even foodborne illnesses[269]. Some of this literature relies on satellite and other remotely-sensed images, such as agricultural productivity assessments and planning[211]. Land cover mapping[240] and socioeconomic status prediction[15] have also been explored. However, these factors are arguably directly visible in satellite data, e.g., to predict socioeconomic status, Ayush et al.[15] search for objects directly in satellite data, such as trucks. Dengue fever prediction in Abdur Rehman et al.[2] is based on identifying features such as standing water locations (mosquito habitat) and roads (human presence). While dengue status is not directly visible, these direct causes are. MND prediction is less direct, as it may depend on disease *and* nearby agriculture, forests, etc.

Possible Causes of MND: The causal mechanisms of MND are complex, but there are multiple factors that likely influence MND, including environmental (e.g., forest presence), epidemiological (e.g., malaria), and socio-economic factors. One of the primary environmental factors studied for its impacts on MND is forests. Generally, research indicates that access to forests may improve dietary diversity. Dietary diversity is an assessment of the range of food groups consumed over a period of time that is typically used as a proxy for sufficient nutrient intake[278], which *is typically measured using survey responses detailing foods consumed.* Forests may directly support dietary diversity, e.g., from bushmeat and wild fruits, provide an additional source of income, e.g., through the sale of forest products, or support crop and livestock production[280]. A study on children's diets across 27 developing countries, including Madagascar, finds that close proximity to forests improved the household prevalence of Vitamin A- and iron-rich foods by 11% and 16%, respectively[244]. Ickowitz et al. [132], one of the most similar studies to ours, analyze dietary diversity, fruit and vegetable consumption, and animal source food consumption in children using satellite data such as tree cover, road location, climate, and urban population information.

As an example of socioeconomic factors, Koppmair et al. [161] show that access (as measured by distance) to food markets in Malawi plays an important role in supporting dietary diversity, particularly for farm households. Markets may directly provide food, and/or may provide additional sources of income for local residents through agricultural and livestock production sales, which can indirectly improve dietary diversity. Agriculture, livestock, and water supply also play an important role in health and nutrition[50]. We further discuss the impact of socioeconomic status on MND in the Appendix[*].

While these methods imply that satellite data can contribute towards predicting MND, dietary diversity depends only on foods consumed, which may be directly observable from satellite imagery (e.g., crops or forests). Biomarkers may involve further subtleties, such as individual characteristics or disease. We use additional features as a result.

[*] https://bit.ly/MND-IAAI2022

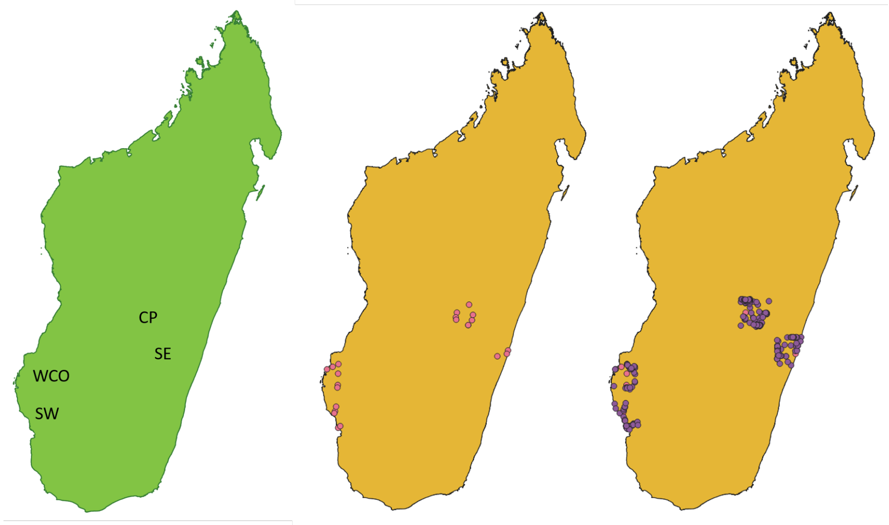


Figure 7.1: Regions studied in Madagascar (left), known (center) and predicted (right) markets in these regions.

7.3 Data Description

Ground Truth Data: Ground truth data were collected by Golden et al.[110] in 2017-2018 in four distinct ecological regions in Madagascar, denoted as the Central Plateau (CP), Southwest (SW), Southeast (SE), and West Coast (WCO) (see Fig. 7.1). CP is at a high elevation, SW is arid, SE is a mid-altitude rainforest, and WCO is seasonally dry.

In this paper, we will focus on the survey responses and biomarker data from blood samples that were collected in Golden et al.[110]. Surveys were provided to individuals in households, small groups, and more. In total, responses were collected from 6292 individuals from 1125 households within 24 communities in CP, SE, SW, and WCO. Biomarker levels from blood draws were also collected from a subset of these individuals. We denote the set of individuals by $p \in \{0, 1, ..., P\}$. Each individual has an underlying MND state, d_p, based on a biomarker level, m, that is thresholded by t, derived from public health literature. Therefore, individual p has $d_p = 1$ if $m < t$ and 0 otherwise. After combining data from blood draws with surveys and household GPS locations, we have 2458

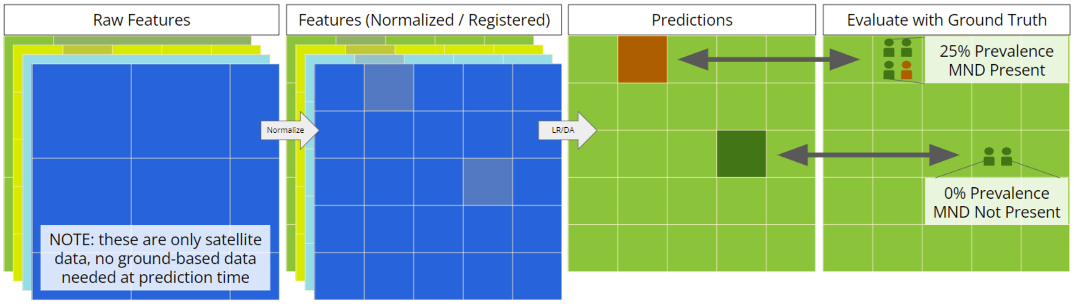

Raw Features
Features (Normalized / Registered)
Predictions
Evaluate with Ground Truth
25% Prevalence
MND Present
Normalize
LR/DA
0% Prevalence
MND Not Present
NOTE: these are only satellite data, no ground-based data needed at prediction time

Figure 7.2: Illustration of using satellite data, which is first normalized and registered, as features to predict MND. Compare to pixel-level labels derived from individual MND statuses. In this illustration, both predictions are correct.

samples.

During this data collection process, Golden et al. [110] followed all procedures to minimize the risk to local populations involved as subjects in the study, as detailed in our approved IRB protocol from the Harvard T.H. Chan School of Public Health (IRB16-0166). This included gaining informed consent for all study-related protocols, including the future cross-referencing of biological data with remotely sensed data products to improve the targeting of public health responses. To briefly summarize this process, a community meeting was held to explain the study using speeches. The research team then visited sampled households to invite individuals to participate. The prospective participants were provided more information if they expressed interest. Furthermore, data are de-identified to limit the risk of breaches of confidentiality, and we follow Harvard IRB protocols to further minimize risk. Gaining informed consent does not automatically alleviate concern of data misuse and inadvertent consequences; nevertheless, we took all necessary precautions to protect human subjects in the study. Please see Golden et al. [110] for further details.

Satellite Data: Based on the causes of MND in Related Work, we select publicly available satellite data, much of which is derived from raw satellite imagery, e.g., using machine learning. We provide a full description of features, including collection time ($\sim$ 2017), in the Appendix, but two we use include livestock population density[255] and weather[200].

Once we collect these features (in the form of images) at the sites of clinical data collection, we resample the images to a uniform resolution of about 25x25 m for one pixel, at a size of 308x308 pixels. This provides us with 23 images total with 86 features each (as image bands). After collecting all satellite data, we normalize each feature to within [0, 1], regardless of whether it was binary, categorical, or continuous. We then do imputation by taking the nearest neighbor if there are any missing data in the feature.

7.4 Problem Description and Aggregation

Given the values from satellite data for a pixel as *input*, our goal is to predict MND presence (classification) or prevalence (regression) in that pixel as the *output*. Ground truth *labels* are derived from biomarkers in blood samples.

Define Grid with Satellite Data: More specifically, we represent the input, i.e., the satellite data, via a multidimensional image array, S. There are 23 S in our dataset, as the ecological regions are large. Therefore, we add an overall image index, $S^{l,r}$, where r represents the current region, and l represents the image index within that region. Each $S^{l,r}$ is indexed by i for rows (y-axis), j for columns (x-axis), and k (z-axis) for features, i.e., the individual satellite data features such as forest cover, weather, and presence of water.

Aggregation to Link Data: To link the two data sources, we rely on locations. Each p (individual, see Data Description) is associated with some g_p, a geographic coordinate. Each $S^{l,r}_{i,j}$ is associated with a set of geographic coordinates, $G^{l,r}_{i,j}$. We may now find the set of individuals, $P^{l,r}_{i,j}$, whose locations fall within each pixel, such that $g_p \in G^{l,r}_{i,j}$. We find their underlying MND states, d_p, to calculate MND prevalence, the percentage of individuals who have MND as defined by biomarker levels. This prevalence, $v_{i,j}^{l,r}$, is our label:

$$v_{i,j}^{l,r} = \frac{\sum_{p \in P_{i,j}^{l,r}} d_p}{|P_{i,j}^{l,r}|}, \tag{7.1}$$

where $|P_{i,j}^{l,r}| = \sum_{p \in P_{i,j}^{l,r}} 1$ is the cardinality of set $P_{i,j}^{l,r}$. We may threshold $v_{i,j}^{l,r}$ for a classification task, or predict the explicit value directly as a regression task. Please see Fig. 7.2 for an illustration. In our dataset, this leads to 300-500 pixel labels, which is only about 0.02% of pixels.

Formally, our goal is to train a *region-specific* ML model $f_\omega^r(\cdot)$ parameterized by ω for each of the 4 ecological regions, where given input training data $S_{i,j}^{l,r}$ in the training set, the model is optimized to minimize the discrepancy between prediction $\hat{v}_{i,j}^{l,r} = f_\omega^r(S_{i,j}^{l,r})$ (see Fig. 7.2) and the *ground truth label* $v_{i,j}^{l,r}$: $\min_\omega \mathbb{E}_{S_{i,j}^{l,r} \in S_{tr}^r} D(\hat{v}_{i,j}^{l,r}, v_{i,j}^{l,r})$ where $D(\hat{v}_{i,j}^{l,r}, v_{i,j}^{l,r})$ could be, e.g., mean squared error (MSE) for regression, or cross-entropy (CE) for classification. $\mathbb{E}_{S_{i,j}^{l,r} \in S_{tr}^r}$ is an expectation taken over all pixels in the training set S_{tr}^r in region r, for each micronutrient. We assume the data are *i.i.d.*

7.5 Prediction Methodology

Market Detection: As discussed in Related Work, the presence of markets is an important factor for MND. We would consequently like to add markets as an extra feature on top of the existing satellite data products. Yet, it is difficult to know where all markets are located in Madagascar. We only know of those specifically mentioned during the focus group surveys conducted in Golden et al.[110].

To add this, we therefore start by comparing the known market locations from the survey data responses with satellite data, and infer that the number of buildings within town clusters and the proximity to roads may be used as predictors of market presence in Madagascar. Specifically, we determine empirically that 20 buildings and one road within about 0.8 km^2 are highly indicative of market presence.

In order to apply these thresholds in an automatic market detection pipeline, we first have to locate roads and buildings. While OpenStreetMap (OSM)[†] provides building and road segmentation data, it is not always complete. This is especially true in our regions of interest. As a result, we train a satellite image-based segmentation model.

For ground truth data to train this segmentation model, we use nearby OSM building labels *where they are more complete*. In particular, for each of the four regions in Madagascar, we automatically identify the closest densely-clustered OSM building labels to the known market locations. These labels are saved to the building segmentation training set, along with high-resolution images from the Google Maps Static API[‡]. For each region, the training dataset contains roughly 100-200 training images and at least 500 corresponding OSM building labels across all images. Each individual image has 600x600 pixels, with a 0.46 m resolution.

For the building segmentation model, we use a U-Net convolutional network[256] with a ResNet-34 encoder pretrained on ImageNet. The U-Net architecture, originally developed for biomedical image segmentation, is commonly used for satellite image segmentation, and is particularly useful for training on smaller training sets such as the sparse OSM building label data. The satellite image training set is augmented with random flips, rotations, and resizes. Binary cross entropy is used as the loss function, and we use the Adam optimizer with a learning rate of 1e-2. The model is trained using a batch size of 16. Results are shown in Fig. 7.1. The building segmentation model and thresholding achieves 0.86 precision in detecting the ground-truth markets from survey data. We include these as features in our data by drawing radii of multiple distances around each market, so that pixels in this layer represent the number of markets within a certain radius. We create these radii masks given healthcare center coordinates[130] as well, bringing us to 90 total features. While we focus on markets here, *this segmentation process could be applied to generate other satellite image-*

[†] www.openstreetmap.org

[‡] developers.google.com/maps/documentation/maps-static

based features that do not already exist, such as custom landcover maps.

K-Medoids-based Feature Selection: It is helpful to have many features, but not all features are necessarily informative. The risk of overfitting when using all 90 features can be large when dealing with limited data. A straightforward idea is to use knowledge from domain experts to select only features that are most important for predicting MND in a particular region. However, this introduces two more issues. First, the feature importance of different regions may vary drastically due to different ecologies. In Madagascar, for example, certain agriculture, such as pulses, are only present and predictive of MND in some regions. It would require a significant amount of manual work to specify the set of important features for each area. Second, the causal mechanisms behind MND are not fully understood. Therefore, it is critical to come up with an automatic feature selection procedure that effectively filters out uninformative features with minimal manual effort.

We start by removing any features that are always 0 throughout the full dataset (i.e., $S_{i,j,k} = 0, \forall i, j$), leading to 69 features. We then use the K-medoids clustering method[232] to group highly correlated features. Each point in our space is a vector of individual pixel values in an image (representing a feature), such that the dimension of the space is the number of pixels. We use Pearson's correlation coefficient as the distance metric between features. Similar to K-means clustering, K-medoids clustering also aims at partitioning the data points (i.e., features) into different clusters. Both minimize the sum of distances between points labeled to be in the same cluster and a point designated to be the center of that cluster. However, K-means uses the central position (centroids) as the designated point, while K-medoids uses a point that actually exists in the set of data points (i.e., an existing satellite data feature). As such, we are able to use the medoid feature to represent the group of correlated features, preserving interpretability.

We post-process the image data, selecting the 300-500 (0.02%) ground truth pixels to form a feature matrix.

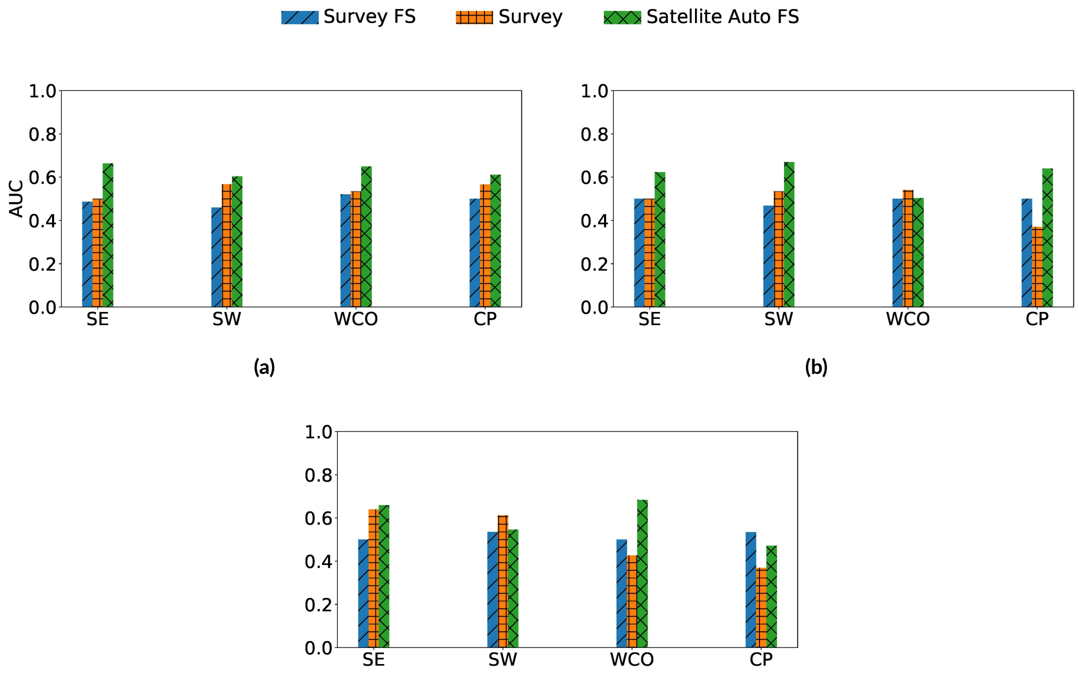
Prediction with Logistic Regression: We first use a simple but effective logistic regression model.

We choose logistic regression as one of the underlying ML models in this paper, due to its following advantages. First, it has fewer weights compared to other models such as deep neural networks, and therefore is less prone to overfitting. This is particularly important given the limited amount of data we have and the high-dimensional feature space. Second, it is interpretable by itself (as shown in experiments, e.g., Fig. 7.5), where the weights ω of different features directly indicate the importance of the features in determining the prediction outcome. Moreover, compared to post-hoc model-free explanation methods such as LIME[251] and SHAP[187], which only provide *instance*-level explanations, the weights of logistic regression models imply feature importance at an *aggregated* level, which we show could provide important insights to public health experts. We primarily focus on region-specific prediction for tailored interpretation and results, but we also train using all regions' training data combined and predict on each regions' test set, which we call Naively Combined.

Prediction with Multi-layer Perceptron and Domain Adaptation: Another strategy to address limited training data is domain adaptation[129], which allows us to use data from all 4 ecological regions as follows: The target domain is the region of Madagascar in which we are making our predictions. The source domains are the other 3 regions, which we would like to use for augmentation. We project all 4 into a domain-invariant latent representation with a single hidden layer (5 neurons) and the loss function:

$$l = \alpha * l_{src} + l_{tgt} + \lambda * l_{transfer} \tag{7.2}$$

where l_{src} and l_{tgt} are the binary cross-entropy loss in the source and target domains. $l_{transfer}$ is the CORAL loss[279] between the source and target domains. α and λ are hyperparameters, and are tuned to be 0.1 and 0.01, respectively, out of $\{0.01, 0.1, 1, 10\}$. Finally, we predict on the target domain test set.



(a)

(b)

(c)

Figure 7.3: Comparison of survey-based (with or without feature selection) and satellite data-based MND prediction by regions. Experimental results. Fig. 7.3a: Iron deficiency, Fig. 7.3b: Vitamin B12 deficiency, Fig. 7.3c: Vitamin A deficiency.

7.6 Results

We present experimental results using 4-fold cross-validation (i.e., data from one region are broken into 4 folds). Due to the limited amount of data, it is impractical to have more folds. We primarily report Area Under the Curve - Receiver Operating Characteristics (AUC-ROC, or AUC in short) to evaluate the MND classification tasks, and discuss recall in the Appendix. Note that we only report the mean AUC values averaged over the 4 folds as the standard deviation becomes trivial for only 4 folds. All data collection and experimentation rely on the default, free resources on Google Colab[§], and training for all 4 folds takes less than 1 minute in general for both logistic regression and

[§] https://colab.research.google.com

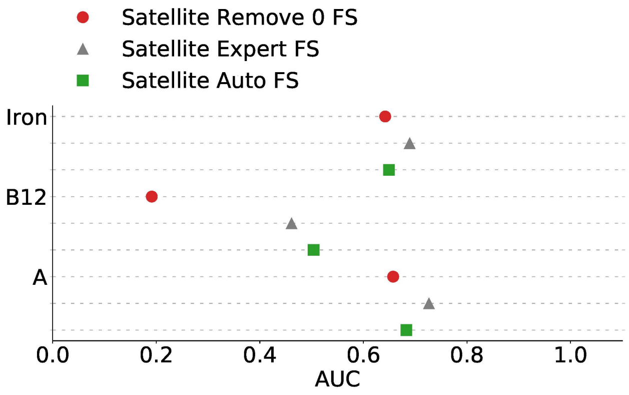


Figure 7.4: Comparison feature selection methods, including removing any features without data, human expert feature selection, and our K-medoids method, all in region WCO.

domain adaptation.

a) Is our prediction accurate? We compare with predictions made by survey data only, as is similar to prior work such as[132]. The results are shown in Fig. 7.3. For survey data, we tested two versions, the original, full amount of data, and a version with one simple level of feature selection. In this case, we selected features which we believed could reasonably be seen or inferred from satellite data. When comparing both survey-based predictions with our satellite data-based predictions, we can see that satellite data-based prediction is better in i) all 4 regions for iron, ii) 3 out of 4 regions for Vitamin B12, and iii) 2 out of 4 regions for Vitamin A. Where it does not outperform survey-based predictions, it performs comparably with significantly lower cost. Across all of the 4 regions and all of the 3 types of nutrients, the AUC value is higher than 0.6 in 10 cases[¶], and is close to 0.5 for the other 2 cases. Meanwhile, the F1 scores of our predictions are on average 0.6 (ranging up to 0.9) and are also comparable to those based on surveys. Satellite data-based regression results are comparable to survey-based regression. Therefore, we consider our predictions accurate.

b) Which features are important for MND prediction? As logistic regression is considered an

[¶] Please note that some of these statistics may slightly fluctuate, e.g., 9 instead of 10 cases sometimes.

Feature Description	Frequency
Chicken population density	9
Cattle population density	8
Net shortwave radiation flux	7
Presence of market within 7.5 km	6
Soil moisture in 100 - 200 cm underground	6
Soil temperature in 10 - 40 cm underground	5
Near surface wind speed	4
Surface pressure	4
Fire (temperature of pixel)	4
Presence of market within 3.75 km	3

Table 7.1: Frequency of each feature appearing in either the top 3 positive or negative coefficients. The 10 (out of 21) features with the highest appearance frequencies are shown.

inherently interpretable model, we focus our analysis on the weights of each variable, particularly those whose absolute values are largest. First, we build an "important features" list. For each region-specific model and each micronutrient (in total $3 \times 4 = 12$ cases), we record the features with the top 3 highest positive weights and negative weights. We aggregated statistics on the number of times that each feature appears in these "important features" lists in Table 7.1. From this, we observe that *market features are very important*, with market presence within 7.5 km with 6 appearances, and within 3.75 km with 3 appearances. We also observe other interesting trends, including that more forest fires are linked to greater rates of Vitamin A and B12 deficiency in the SE region (rainforest), but not in other regions that are less reliant on forest products, which may be a useful insight for public health experts. Fig. 7.5 illustrates this pattern for Vitamin A in SE.

c) How does the automatic feature selection perform? To evaluate the performance of automatic feature selection (FS), we compare with two baselines. First, we consider the case where there is no feature selection apart from removing features which are completely zero (i.e., no data) (Satellite Remove 0 FS). We also compare with expert feature selection, in which a public health expert examines

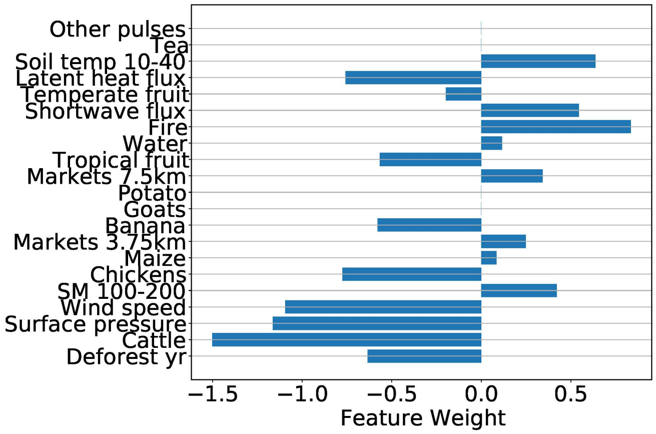


Figure 7.5: Logistic regression weights (x-axis) for Vitamin A, region SE. Positive numbers mean positive correlation with MND. Medoid feature names provided (SM: soil moisture).

the features we propose, and groups them based on their knowledge[‖]. They also select a representative feature for each of their groups (Satellite Expert FS). Finally, we consider the performance of our correlation and K-medoids-based algorithm (Satellite Auto FS). We show results for one of the regions (WCO) due to space limitation, but trends in other regions are similar. We can see that both Expert FS and Auto FS are better than the case where no FS is used, especially for Vitamin B12. In all three cases, Auto FS always performs comparably to Expert FS, as it does in other examples that are not included here, but Auto FS is more scalable.

We also compare the groups that are found by Auto FS and Expert FS. Very interestingly, we find that in the two methods, 8 out of 21 group centers overlap: banana, cattle, chicken, goat, maize, presence of markets within 7.5 km, surface pressure, and wind speed. This shows that our method is choosing features deemed important by a human expert as well. The above results well demonstrate that our proposed automatic feature selection method is an effective while scalable alternative to expert feature selection.

d) How do different prediction paradigms compare? We compare the region-specific logistic re-

[‖] Expert chose 21, which led us to select $K = 21$

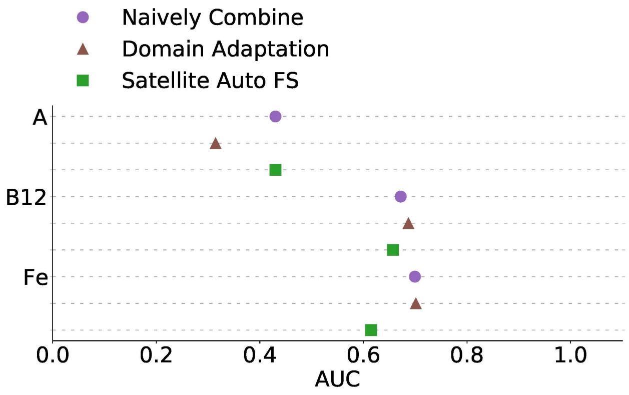


Figure 7.6: Comparing AUC of a logistic regression model trained by naively combining training data from all regions, a multi-layer perceptron with domain adaptation, and a region-specific logistic regression model, all in CP.

gression models (Satellite Auto FS), the logistic regression model version that combines training data from all of the regions (Naively Combined), and multi-layer perceptron with domain adaptation (Domain Adaptation). We present results from region CP. Here, and overall, we find that Vitamin B12 and Iron achieve better performance using Domain Adaptation, while Vitamin A achieves better performance using the logistic regression-based Satellite Auto FS or Naively Combined. This may be because each micronutrient differs slightly in its relevant factors, and factors may vary regionally (e.g., some regions are forested). Clearly, each method works well with limited amounts of data, but we acknowledge the tradeoffs in interpretability, and a potential lack of robustness in the model due to limited samples.

7.7 Conclusion and Discussion

In conclusion, satellite data are viable to use for MND prediction at a public health scale. We presented a system relying on the aggregation of individual MND states over geographic regions, a search for relevant features, such as markets, automatic feature selection, which performs comparably to human expert feature selection, and domain adaptation and logistic regression prediction models. This system worked well even with limited ground truth biomarker data.

Deployment Considerations While our system has not yet been deployed, we would like to emphasize several deployment considerations. This methodology would not replace surveys and blood samples collected among communities. Rather, we believe it should be used to cover gaps in that data collection, e.g., where data could not be collected, or in between collections. To do this, public health officials, policymakers, healthcare workers, or individuals can load publicly available, current satellite data and apply the existing model, without any survey or blood sample data. We can then update these models when another data collection occurs. This also applies for deployment in other countries. We plan to develop a web application to load satellite data at the desired time and location, and the current proposed model, to provide predictions. We plan to iterate on this with potential users, including officials from Catholic Relief Services, Médecins Sans Frontières, and the Ministry of Health in Madagascar. In the meantime, code and satellite data are available[**], while ground truth data are withheld for privacy.

Future Work: We began preliminary experiments into sparse segmentation and spatial aggregation to further include spatial patterns in the prediction step, but they require further refinement before deployment. We also encourage the use of custom features, as illustrated with markets. Most importantly, we believe there is ample room for further research, both in this domain and others with sparse data, and a great deal of promise for broad application to inform future public health interventions.

[**] https://github.com/exb7900/mnd-iaai2022

8

Envisioning Communities: A Participatory Approach Towards AI for Social Good

8.1 Introduction

Artificial intelligence (AI) for social good, hereafter AI4SG, has received growing attention across academia and industry. Countless research groups, workshops, initiatives, and industry efforts tout

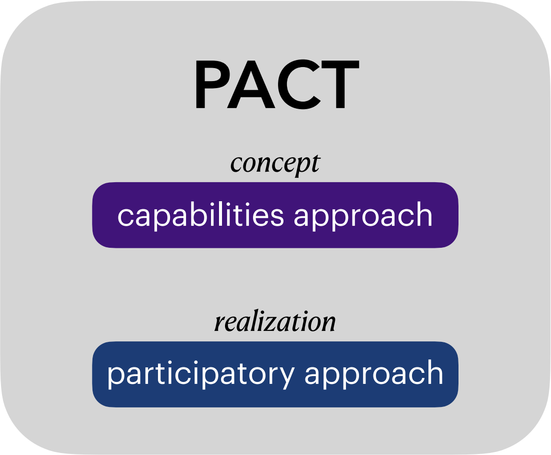


Figure 8.1: The framework we propose, Participatory Approach to enable Capabilities in communiTies (PACT), melds the capabilities approach (see Figure 8.3) with a participatory approach (see Figures 8.4 and 8.5) to center the needs of communities in AI research projects.

programs to advance computing and AI "for social good." Work in domains from healthcare to conservation has been brought into this category[269,309,169]. We, the authors, ourselves have endeavored towards AI4SG work as computer science and philosophy researchers.

Despite the rapidly growing popularity of AI4SG, social good has a nebulous definition in the computing world and elsewhere[114], making it unclear at times what work ought to be considered social good. For example, can a COVID-19 contact tracing app be considered to fall within this lauded category, even with privacy risks? Recent work has begun to dive into this question of defining AI4SG[95,191,94]; we will explore these efforts in more depth in Section 8.2.

However, we point out that context is critical, so no single tractable set of rules can determine whether a project is "for social good." Instead, whether a project may bring about social good must be determined by those who live within the context of the system itself; that is, the community that it will affect. This point echoes recent calls for decolonial and power-shifting approaches to AI that focus on elevating traditionally marginalized populations[205,144,174,201,31,312].

The community-centered, context-specific conception of social good that we propose raises its own questions, such as how to reconcile multiple viewpoints. We therefore address these concerns with an integrated framework, called the **P**articipatory **A**pproach to enable **C**apabilities in communi**T**ies, or PACT, that allows researchers to assess "goodness" across different stakeholder groups and different projects. We illustrate PACT in Figure 8.1. As part of this framework, we first suggest ethical guidelines rooted in capability theory to guide such evaluations[264,223]. We reject a view that favors courses of action solely for their aggregate net benefits, regardless of how they are distributed and what resulting injustices arise. Such an additive accounting may easily err toward favoring the values and interests of majorities, excluding traditionally underrepresented community members from the design process altogether.

Instead, we employ the *capabilities approach*, designed to measure human development by focusing on the substantive liberties that individuals and communities enjoy, to lead the kind of lives they have reason to value[264]. A capability-focused approach to social good is aimed at increasing opportunities for people to achieve combinations of "functionings"—that is, combinations of things they may find valuable doing or being. While common assessment methods might overlook new or existing social inequalities if some measure of "utility" is increased high enough in aggregate, our approach would only define an endeavor as contributing to social good if it takes concrete steps toward empowering all members of the affected community to each enjoy the substantive liberties to function in the ways they have reason to value.

We then propose to enact this conception of social good with a *participatory approach* that involves community members in "a process of investigating, understanding, reflecting upon, establishing, developing, and supporting mutual learning between multiple participants"[273], in order for the community itself to define what those substantive liberties and functionings should be. In other words, communities define social good in their context.

Our contributions are therefore (i) arguing that the capabilities approach is a worthy candidate for conceptualizing social good, especially in diverse-stakeholder settings (Section 8.3), (ii) highlighting the role that AI can play in expanding and equalizing capabilities (Section 8.4), (iii) explaining how a participatory approach is best served to identify desired capabilities (Section 8.5), and (iv) presenting and discussing our proposed guiding principles of a participatory approach (Section 8.6). These contributions come together to form PACT.

8.2 Growing Criticisms of AI for Social Good

As a technical field whose interactions on social-facing problems are young but monumental in impact, the field of AI has yet to fully develop a moral compass. On the whole, the subfield of AI for social good is not an exception.

We highlight criticisms that have arisen against AI4SG research, which serve as a call to action to reform the field. Later, we argue that a participatory approach rooted in enabling capabilities will provide needed direction to the field by letting the affected communities—particularly those who are most vulnerable—be the guide.

Recent years have seen calls for AI and computational researchers to more closely engage with the ethical implications of their work. Green[113] implores researchers to view themselves and their work through a political lens, asking not just how the systems they build will impact society, but also how even the problems and methods they choose to explore (and *not* explore) serve to normalize the types of research that ought to be done. Latonero[170] offers a critical view of technosolutionism as it has recently manifested in AI4SG efforts emerging from industry, such as Intel's TrailGuard AI that detects poachers in camera trap images, which have the potential to individually identify a person. Latonero argues that while companies may have good intentions, they often lack the ability to gain the expertise and local context required to tackle complex social issues. In a related spirit, Blumenstock[33] urges researchers not to forget "the people behind the numbers" when developing data-driven solutions, especially in development contexts. De-Arteaga et al.[81] focus on a specific subset of AI4SG dubbed machine learning for development (ML4D) and similarly express the importance of considering local context to ensure that researcher and stakeholder goals are aligned.

Also manifesting recently are meta-critiques of AI4SG specifically, which contend that the subfield is vaguely defined, to troubling implications. Moore[206] focuses specifically on how the choice of the word "good" can serve to distract from potentially negative consequences of the development of certain technologies, retorting that AI4SG should be re-branded as AI for "not bad". Malliaraki[196] argues that AI4SG's imprecise definition hurts its ability to progress as a discipline, since the lack of clarity around what values are held or what progress is being made hinders the ability of the field to establish specific expertise. Green[114] points out that AI4SG is sufficiently vague to encompass projects aimed at police reform as well as predictive policing, and therefore simply lacks meaning. Green also argues that AI4SG's orientation toward "good" biases researchers toward incremental technological improvements to existing systems and away from larger reformative efforts that could be better.

Others have gone further, calling for reforms in the field to say that "good" AI should seek to shift power to the traditionally disadvantaged. Mohamed et al.[205] put forth a decolonial view of AI that suggests that AI systems should be built specifically to dismantle traditional forms of colonial oppression. They provide examples of how AI can perpetuate colonialism in the digital age, such as through algorithmic exploitation via Mechanical Turk–style "ghost work"[111] or through algorithmic dispossesion in which disadvantaged communities are designed *for* without being allowed a seat at the design table. They also offer three "tactics" for moving toward decolonial AI, namely: (1) a critical technical practice to analyze whether systems promote fairness, diversity, safety, and mechanisms of anticolonial resistance; (2) reciprocal engagements that engender co-design between affected communities and researchers; and (3) a change in attitude from benevolence to solidarity, that again necessitates active engagement with communities and grassroots organizations. In a similar spirit, Kalluri[144] calls for researchers to critically analyze the power structures of the systems they design for and consider pursuing projects that empower the people they are intended to help. For example, researchers may seek to empower those represented in the data that enables the system, rather than the decision-maker who is privileged with a wealth of data. This could be accomplished by designing systems for users to audit or demand recourse based on an AI system's decision[144].

In tandem with these criticisms, there have been corresponding efforts to define AI4SG. Floridi et al.[94] provide a report of an early initiative to develop guidelines in support of AI for good. Therein, they highlight risks and opportunities for such systems, outline core ethical principles to be considered, and offer several recommendations for how AI efforts can be given a "firm foundation" in support of social good. Floridi et al.[95] later expand on this work by proposing a three-part account, which includes a definition, a set of guiding principles, and a set of "essential factors for success." They define AI4SG as "the design, development, and deployment of AI systems in ways that (i) prevent, mitigate, or resolve problems adversely affecting human life and/or the well-being of the natural world, and/or (ii) enable socially preferable and/or environmentally sustainable developments." Note that this disjunctive definition captures a broad spectrum of projects with widely diverging outcomes, leaving open still the possible critiques discussed above. However, their principles and essential factors contribute to establishing rough guidelines for projects in the field, as others have done[293,191] in lieu of seeking a definition.

8.3 The Capabilities Approach

Clearly, AI for social good is in need of a stricter set of ethical criteria and stronger guidance. To understand what steps take us in the direction of a more equitable, just, and fair society, we must find the right conceptual tools and frameworks. Utilitarian ethics, a widely referenced framework, adopts the aggregation of utility (broadly understood) as the sole standard to determine the moral value of an action[292,254].

According to classic utilitarianism, the right action to take in any given context is whichever action maximizes utility for society. This brand of utilitarianism seems particularly attractive in science-oriented circles, due to its claim that a moral decision can be made by simply maximizing an objective function. For example, in the social choice literature, Bogomolnaia et al.[35] argue for the use of utilitarianism as "it is efficient, strategyproof and treats equally agents and outcomes." However, the apparent transparency of the utilitarian argument obscures its major shortcomings. First, like the problem of class imbalance in machine learning, a maximizing strategy will bias its objective towards the majority class. Hence, effects on minority and marginalized groups may be easily overlooked. Second, which course of action can maximize utility for society overall cannot be determined on simple cost–benefit analyses. Any such procedure involves serious trade-offs, and a moral decision requires that we acknowledge tensions and navigate them responsibly.

To take a recent example, the development of digital contact tracing apps in the wake of the COVID-19 pandemic represented a massive potential benefit for public health, yet posed a serious threat to privacy rights, and all the values that privacy protects—including freedom of thought and expression, personal autonomy, and democratic legitimacy. Which course of action will maximize utility? It is not just controversial. What verdict is reached will necessarily depend on what fundamental choices and liberties individuals value most, and which they are willing to forgo. Furthermore, for some groups the losses may be more significant than for others, and this fact requires due consideration.

A related, though distinct, standard might put aside the ambition of maximizing good and adopt some threshold condition instead. As long as it generates a net improvement over the *status quo*, someone might say, an action can be said to promote social and public good. This strategy raises three additional problems: first, utility aggregates are insensitive to allocation and distribution; second, they are not sensitive to individual liberties and fundamental rights; and third, to the extent that they consider stakeholders' preferences, they do not take into account the idea that individuals' preferences depend on their circumstances and in better circumstances could have different preferences[264]. Why should this matter? *Because attempts to increase utility may be too easily signed off as progress while inflicting serious harm and contributing to the entrenchment of existing inequality.*

For this reason, we believe that respect for liberty, fairness of distribution, and sensitivity to interpersonal variations in utility functions should serve as moral constraints, to channel utility increments towards more valuable social outcomes. Operating within these constraints suggests shifting the focal point from utility to a different standard. In the spirit of enabling capabilities, we suggest that such standard should be based on an understanding of the kinds of lives individuals have reason to value, how the distribution of resources in combination with environmental factors may create (or eliminate) opportunities for them to actualize these lives, and designing projects that create fertile conditions for these opportunities to be secured.

This reorientation of the moral assessment framework is in line with a view of AI as a tool to shift power to individuals and groups in situations of disadvantage, as advocated by Kalluri[144]. What this shift means, in operational terms, is not yet fully elucidated. We therefore hope to contribute to this endeavor by exploring conceptual tools that may guide developers in the process, providing intermediate landmarks. We consider "capabilities" to be suitable candidates for this task, as they constitute a step in the direction of empowering disadvantaged individuals to decide what, in their view, that shift should consist of.

The capabilities approach, which stands at the intersection of ethics and development economics, was originally proposed by Amartya Sen and Martha Nussbaum[264,224]. The approach focuses on the notion that all human beings should have a set of substantive liberties that allow them to function in society in the ways they choose[224]. These substantive liberties include both freedom from external obstruction and the external conditions conducive to such functioning. The *capability set* includes all of the foundational abilities a human being ought to have to flourish in society, such as, freedom of speech, thought and association; bodily health and integrity; and control over one's po-

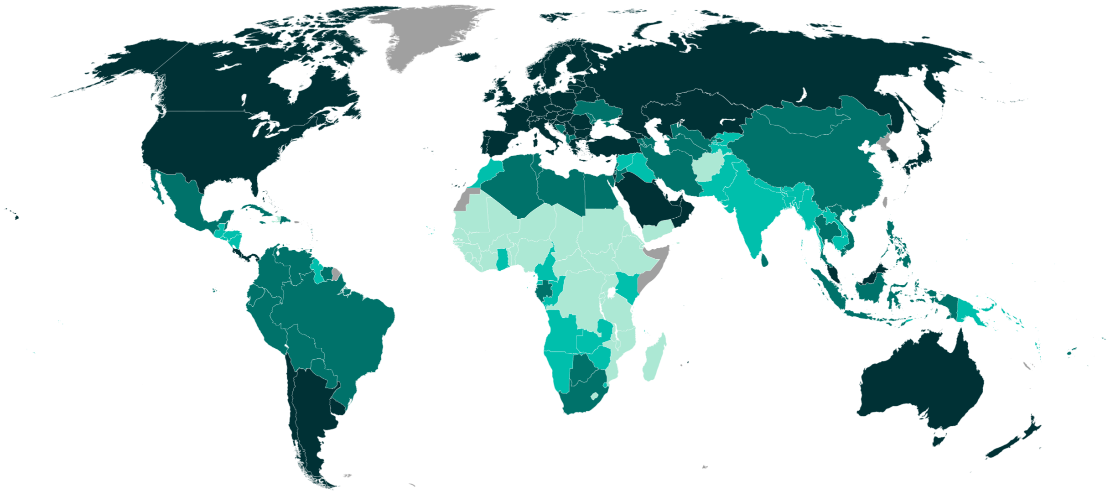


Figure 8.2: The United Nations Human Development Index (HDI) was inspired by the capabilities approach, which serves as the foundation of the participatory approach that we propose. The above map from Wikimedia Commons shows the HDI rankings of countries around the world, where darker color indicates a higher index.

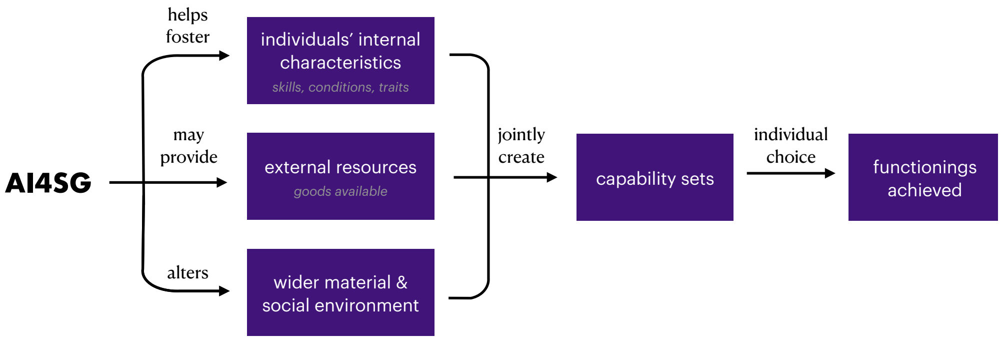


Figure 8.3: AI for social good research projects can expand and equalize capabilities by fostering individuals' internal characteristics, providing external resources, or altering the wider material and social environment, thus creating capability sets to improve social welfare and fight injustice.

litical and material environment; among others. The capabilities approach has inspired the creation of the United Nations Human Development Index (Figure 8.2), marking a shift away from utilitarian evaluations such as gross domestic product to people-centered policies that prioritize substantive capabilities[299].

A capability-oriented AI does not rest content with increasing aggregate utility and respecting legal rights. It asks itself what it is doing, and how it interacts with existing social realities, to enhance or reduce the opportunities of the most vulnerable members of society to pursue the lives they have reason to value. It asks whether the social, political, and economic environment that it contributes to create deprives individuals of those capabilities or fosters the conditions in which they may enjoy equal substantive liberties to flourish. In this framework, a measure of social good is how much a given project contributes to bolster the enjoyment of substantive liberties, especially by members of marginalized groups. This kind of measure is respectful of individual liberties, sensitive to questions of distribution, and responsive to interpersonal variations in utility.

Before discussing the relation between AI and capabilities, we would like to dispel a potential objection to our framework. The capabilities approach has been criticized for not paying sufficient attention to groups, institutions, and social structures, thus rendering itself unable to account for power imbalances and dynamics[126,158]. This characteristic would make the approach unappealing for a field that seeks to address injustices in the distribution of power. However, the approach does indeed place substantial emphasis on conceptualizing the factors that affect (particularly those which may enhance) individuals' power. Its emphasis on substantial opportunities is itself a way of conceptualizing what individuals have the power to do or to be. These powers are determined in large part by the broader social environment, which includes group membership, inter-group relations, institutions, and social practices. Furthermore, the approach explicitly focuses on enhancing the power of the most vulnerable members of society. In fact, the creation and enhancement of capabilities constitute intermediate steps on the way towards shifting power, promoting the welfare and broadening the liberties of those who have been historically deprived. For this reason, it provides a fruitful framework to assess genuine moral progress and a promising tool for AI projects seeking to promote social good.

Light discussion of the capabilities approach has begun inside the AI community. Moore[206] references the capabilities approach to call for greater accountability and individual control of private data. Coeckelbergh[70] proposes taking a capabilities approach to health care, particularly when using AI assistance to replace human care, but focuses on Nussbaum's proposed set of capabilities rather than eliciting desired capabilities from affected patients. We significantly build on these claims by discussing the potential of AI to enhance capabilities and argue that the capabilities approach goes hand-in-hand with a participatory approach. Two papers specifically ground their work in the capabilities approach; we highlight these as exemplars for future work in AI4SG. Thinyane & Bhat[291] invoke the capabilities approach, specifically to empower the agency of marginalized groups, to motivate their development of a mobile app to identify victims of human trafficking in Thai fisheries. Kumar et al.[167] conduct an empirical study of women's health in India through the lens of Nussbaum's central human capabilities.

8.4 AI and Capabilities

The capabilities approach may serve AI researchers and developers to assess the potential impact of projects and interventions along two dimensions of social progress: capability expansion and distribution. This section argues that, when they interact productively with other social factors, AI projects can contribute to equalizing and enhancing capabilities—and that *therein* lies their potential to bring about social good. For instance, an AI-based system Visual Question Answering System can enhance the capabilities of the visually impaired to access relevant information about their environment[155].

When considering equalizing and enhancing capabilities, it is important to notice that whether an individual or group enjoys a set of capabilities is not solely a matter of having certain personal characteristics or being free from external obstruction. External conditions need also be conducive to enable individuals' choices among the set of alternatives that constitutes the capability set[224]. This is why these liberties are described as *substantial*, as opposed to formal or negative. Hence, capabilities, or substantive liberties, are composed of (1) individuals' personal characteristics (including skills, conditions, and traits), (2) external resources they have access to, and (3) the configuration of their wider material and social environment[142].

Prior work at the intersection of technology and capabilities has addressed the potential of technology to empower users, by enhancing their capabilities and choice sets. Johnstone[142] proposes the capabilities approach as a key framework for computer ethics, suggesting that technological objects may serve as tools or external resources that enhance individuals' range of potential action and choice[71,156]. This is an important role that AI-based technologies may play in enhancing capabilities: providing tools for geographic navigation, efficient communications, accurate health diagnoses, and climate forecasts. Technology may also foster the development of new skills, abilities, and traits that broaden an individual's choice sets. We diagram in Figure 8.3 the relationship between an AI4SG project on an individual's characteristics, resources, and environments, and thus the potential of AI to alter (both positively and negatively) capability sets.

Technological objects may also become part of social structures, forming networks of interdependencies with people, groups, and other artifacts[228,316]. When focusing on AI-based technologies, it is crucial to also acknowledge their potential to affect the social and material environment, which may render it more or less conducive to secure capability sets for individuals and communities. For example, AI-based predictive policing may increase the presence of law enforcement agents in specific areas, which may in turn affect the capability sets of community members residing in those areas. If this presence increases security from, say, armed robbery, community members may enjoy some more liberties than they previously did. And yet, if this acts to reinforce the disparate impact of law enforcement on members of vulnerable communities, along with other inequalities, then it may not only diminish the substantive liberties of those individuals impacted, but also alter the way in which such liberties are distributed across the population.

Assessing this kind of trade-off ought to be a crucial step in evaluating the ability of a particular project to promote social good. This assessment must be aligned with the kinds of choices and opportunities community members have reasons to pursue[266]. A project should not be granted the moral standing of being "for social good" if it leads to localized utility increments, at the expense of reducing the ability of members of the larger community to choose the lives they value. Most importantly, this assessment must be made by community members themselves, as the following section argues.

8.5 Community Participation

As Ismail & Kumar[136] contend, communities should ultimately be the ones to decide whether and how they would like to use AI. If the former condition is met and the community agrees that an AI solution may be relevant and useful, the latter requires the inclusive design of an AI system through a close and continued relationship between the AI researcher and those impacted.

This close partnership is particularly important as the gross effects of AI-based social interventions on communities' capabilities are unlikely to be exclusively positive. Tradeoffs may be forced on designers and stakeholders; some are likely to be intergroup, and some, intragroup. For this reason, it is only consistent with the proposed approach that designers consult stakeholders from all groups on their preferred ways to navigate such tradeoffs. In other words, if PACT is focused on creating capabilities, it must enable impacted individuals to have a say on what alternatives are opened (or closed) to them.

Moreover, AI projects that incorporate stakeholders' choices into their design process may contribute to create what Wolff & De-Shalit[318] refer to as "fertile functionings." That is, functionings that, when secured, are likely to secure other functionings. Fertile functionings include, though are not limited to, the ability to work and the ability to have social affiliations. These are the kinds of functionings that either enable other functionings (e.g. control over one's environment) or reduce the risk associated with them.

Projects in AI that create propitious environments and enable individuals to make decisions over the capability sets they value in turn give those individuals the capability to function in a way that leads to the creation of other capabilities. If this kind of participatory space is offered to those who are the most vulnerable, AI could plausibly act against disadvantage, and contribute to shifting the distribution of power. In this way, the PACT framework is aligned with the principles of Design Justice in prioritizing the voices of affected communities and viewing positive change as a result of an accountable collaborative process[73].

The PACT framework is also committed to the notion that, when applying capabilities to AI, we must include all stakeholders, especially vulnerable members of society. But how do we accomplish this concretely, and how does this affect the different steps in an AI4SG research project? In the following section, we will first introduce suggested mechanisms of a participatory approach rooted in capabilities, and then discuss how these come into play in (1) determining which capability sets to pursue at the beginning of a project, and (2) evaluating the success of an AI4SG system in terms of its impact, particularly on groups' sets of substantive liberties.

The approach we propose is diagrammed in Figure 8.4, which embeds community participation into each stage. By beginning first with defining capability sets, we resist the temptation to immediately apply established tools to address ingrained social issues, which would only restrict the possibilities for change[184]. These participatory approaches constitute bottom-up approaches for embedding value into AI systems, by learning values from human interaction and engagement

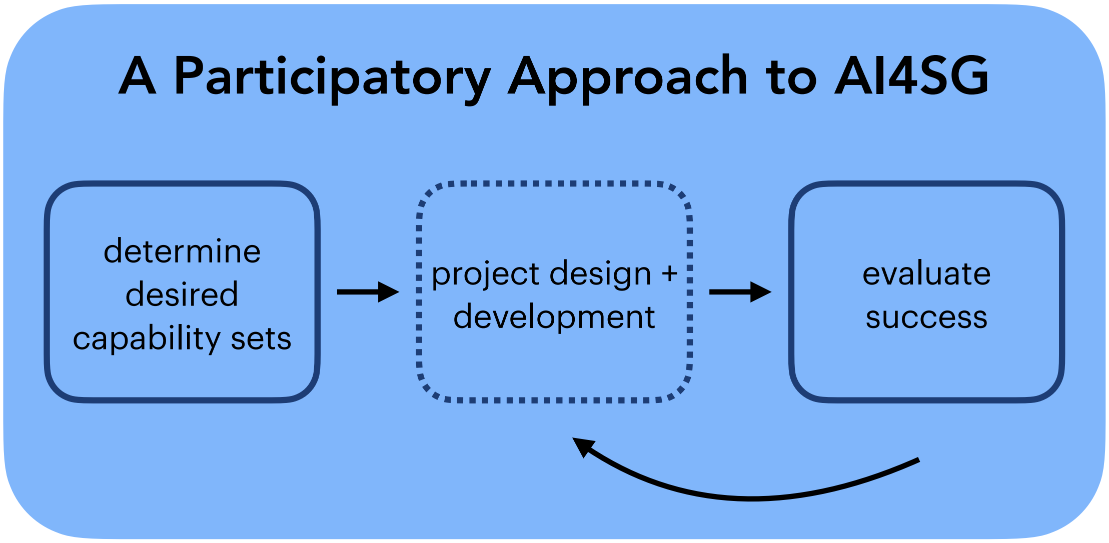

A Participatory Approach to AI4SG
determine desired capability sets
project design + development
evaluate success

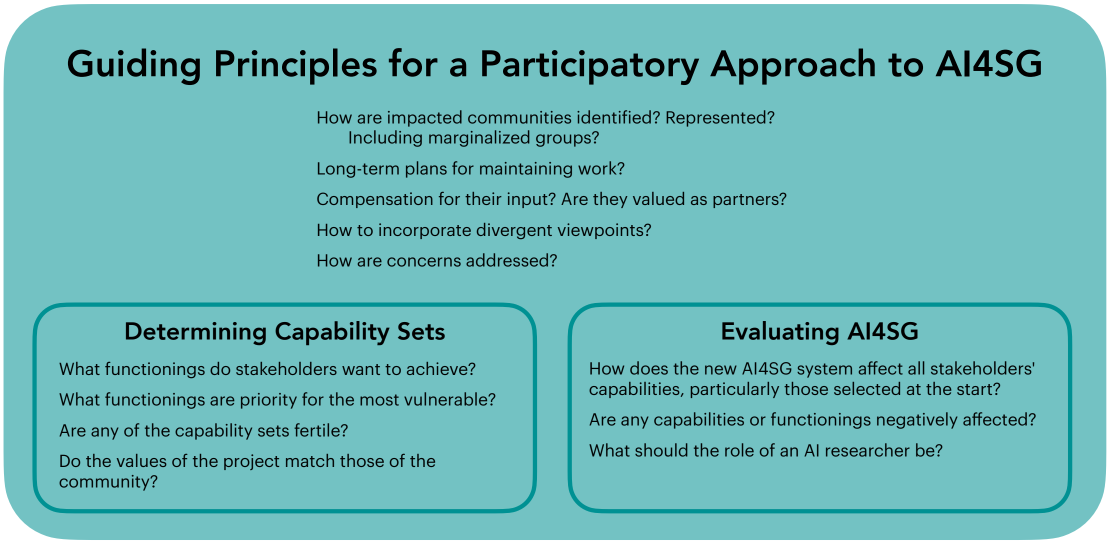
Figure 8.4: Our proposed approach to AI4SG projects, where the key is that stakeholder participation is centered throughout. We do not explicitly comment on the design and development of the project other than calling for the inclusion of community participation and iterative evaluation of success, in terms of whether the project realized the desired capability sets.

rather than being decided through the centralized and non-representative lens of AI researchers[178].

8.6 Guiding Principles for a Participatory Approach

To direct the participatory approach we propose, we provide the following set of guiding questions, outlined in Figure 8.5 and elaborated on below. These questions are specifically worded to avoid binary answers of yes or no. When vaguely stated requirements are put forth as list items to check off, there is a risk of offering a cheap seal of approval on projects that aren't sufficiently assessed in terms of their ethical and societal implications. Instead, we expect that in nearly all cases, none of these questions will be fully resolved. Thus, these questions are meant to serve as a constant navigation guide to help researchers regularly and critically reassess their intended work.

We begin our discussion on how a participatory approach to AI4SG would look with our first guiding question:



Guiding Principles for a Participatory Approach to AI4SG

How are impacted communities identified? Represented? Including marginalized groups?

Long-term plans for maintaining work?

Compensation for their input? Are they valued as partners?

How to incorporate divergent viewpoints?

How are concerns addressed?

Determining Capability Sets

What functionings do stakeholders want to achieve?

What functionings are priority for the most vulnerable?

Are any of the capability sets fertile?

Do the values of the project match those of the community?

Evaluating AI4SG

How does the new AI4SG system affect all stakeholders' capabilities, particularly those selected at the start?

Are any capabilities or functionings negatively affected?

What should the role of an AI researcher be?

Figure 8.5: Guiding principles for the participatory approach we propose. The key message is to center the community's desired capabilities throughout, and to ensure that the goals of the project are properly aligned to do so without negatively impacting other capabilities.

How are impacted communities identified and how are they represented in this process? Who represents historically marginalized groups?

This question is one of the most important and potentially difficult. As such, we encourage readers to research their domain of interest and seek partners as an important first step, but offer some advice from our experience. We often consider seeking partners at non-profits, community-based organizations, and/or (non-)governmental organizations (NGOs), and have even had some experience with non-profits or NGOs finding us. Finding these groups as an AI researcher may be done with the help of an institution's experts on partnerships (e.g., university or company officials focused on forming external partnerships), prior collaborators or acquaintances, discussions with peers within an institution in different departments (e.g., public health or ecology researchers), or "cold emails." Young et al. [338] provide further guiding questions for finding these partners, particularly in their companion guide[193]. In any case, some vetting and research is an important step.

AI researchers may also co-organize workshops, such as a pair of "AI vs. tuberculosis" workshops held by some of the authors in Mumbai, India which brought together domain experts, non-profits, state and local health officials, industry experts, and researchers, identified by Mumbai-based collaborators who specialize in building relationships across health-focused organizations in India. The varied backgrounds of the stakeholders at these workshops was conducive to quickly identifying areas of greatest need across the spectrum of tuberculosis care in India, many of which would not have been considered by AI researchers alone. Further, these group forums sparked conversations between stakeholders that often would not otherwise communicate, but that led to initial ideas for solutions. This highlights that such approaches are often most successful when AI researchers move out of their usual spheres and into the venues of the domains that their projects impact.

At the inception of a project, the designers and initial partners should together identify all relevant stakeholders that will be impacted by the proposed project, bringing them on as partners or members of a (possibly informal) advisory panel. Every effort should be made to include at least one member from each impacted group. If stakeholders were initially omitted and are later identified, they should then be added. Initial consultation with panel members and partners should be carried out in a way that facilitates open and candid discussion to ensure all voices are represented and accounted for during the project design phase[338]. We additionally propose that during the lifetime of the project, the panel and partners should, if possible, be available for ongoing consultation during mutually agreed upon checkpoints to ensure the project continues to align with stakeholder values. Similar practices have been suggested by Perrault et al.[237], who advocate for close, long-term collaborations with domain experts. Likewise, the Design Justice network pursues non-exploitative and community-led design processes, in which designers serve more as facilitators than experts. This framework is modeled on the idea of a horizontal relation in which affected communities are given the role of domain experts, given their lived experience and direct understanding of projects' impact[73].

How are plans laid out for maintaining and sustaining this work in the long-term, and how would the *partnership be ended?*

We believe partnership norms should be established at the beginning of the partnership, including ensuring that the community representatives and experiential experts have the power to end the project and partnership, in addition to the designers and other stakeholders. As noted by Madaio et al. [191], this should also necessitate an internal discussion amongst the research team to decide what, if any, criteria would be grounds for such a cessation, e.g., if it becomes clear that the problem at hand requires methods in which the research team does not have sufficient expertise. Discussions with the broader partners should also lay out the expected time scale and potential benefits and risks for all involved.

Once the relationship is underway, it should be maintained as discussed, with regular communication. When a project reaches the point where a potential deployment is possible, we advocate for an incremental deployment, such as the example of Germany's *teststrecken* for self-driving cars, which sets constraints for autonomy, then expands these constraints once certain criteria are met[95].

What kind of compensation are stakeholders receiving for their time and input? Does that compensation respect them as partners?

Stakeholders must be respected as experiential or domain experts and be considered as partners. They must be compensated in some way, whether monetarily or otherwise. We advocate for compensation to avoid establishing an approach made under the guise of participatory AI that ends up being extractive[231]. One notable drawback to a participatory approach is the potential for these community interactions to become exploitative if not done with intention, compensation, or long-term interactions[276]. Collaborators who have significantly influenced the design or implementation of a project ought to be recognized as coauthors or otherwise acknowledged in resulting publications, presentations, media coverage, etc.

How can we understand and incorporate viewpoints from many groups of stakeholders?

Surveys or voting may seem a natural choice. However, a simple voting mechanism is risky, as it may find that a majority of the community favors, for example, building a heavily polluting factory near a river, while the impacted community living at the proposed site would object that this factory would severely degrade their quality of life. These concerns from the marginalized group must be given due weight. This emphasis on the welfare of marginalized groups is based on the premise that the evaluation of human capabilities must consider *individual capabilities*[223]. In short, all individual capabilities are valuable as ends in their own right; they should never be considered means to someone else's rights or welfare. We must, therefore, guard against depriving individuals of basic entitlements as a means to enhancing overall welfare.

Hence, we endorse a deliberative approach, which aims to uncover "overlapping consensus"[246]. This approach is based on the expectation that, in allowing diverse worldviews, value systems, and preference sets to engage in conversation, with appropriate moderation and technical tools, social groups may find a core set of decisions that all participants can reasonably agree with. Deliberative approaches to democracy have been operationalized by programs such as vTaiwan, which uses digital technology to inform policy by building consensus through civic engagement and dialogue[128]. vTaiwan uses the consensus-oriented voting platform Pol.is which seeks to identify a set of common values upon which to shape legislation, rather than arbitrating between polarized sides[289]. Similarly, OPPi serves as a platform for consensus building through bottom-up crowd-sourcing[179]. This tool is tailored for opinion sharing and seeking, oriented towards finding common ground among stakeholders.

An alternative to fully public deliberative processes are targeted efforts such as Diverse Voices panels[338] which focus on including traditionally marginalized groups in existing policy-making processes. Specifically, they advocate for informal elicitation sessions with partners, asking questions such as, "What do you do currently? What would support your work?"[151]. They also suggest considering whether tech should be used in the first place and highlight that a key challenge of participatory design is to determine a course of action if multiple participants disagree. We add that it remains a challenge to assemble these panels.

Simonsen & Robertson[273] provide multiple strategies as well, particularly with the goal of coming to a common language and fostering communication between groups of experts from different backgrounds. They suggest strategies to invite discussion, such as games, acting out design proposals, storytelling, group brainstorms of what a perfect utopian solution would look like, participatory prototyping, and probing. In the case of probing, for example, one strategy was to provide participants with a cultural probe kit, consisting of items such as a diary and a camera, in order to understand people's reactions to their environments[101].

Further, several fields in artificial intelligence have devoted a great deal of thought to learning and aggregating preferences: preference elicitation[60], which learns agent utility functions; computational social choice[47], which deals with truthful agents; and mechanism design[271], which deals with strategic agents. As an example, Kahng et al.[143] form what they call a "virtual democracy" for a food rescue program. They collect data about preferences on which food pantry should receive food, then use these preferences to create and aggregate the preferences of virtual voters. The goal is to make ethical decisions without needing to reach out to stakeholders each time a food distribution decision needs to be made.

These fields have highlighted theoretical limitations in preference aggregation, such as Arrow's impossibility theorem[13] which reveals limitations in the ability to aggregate preferences with as few as three voters, but these results do not necessarily inhibit us from designing good systems in practice[265,198]. Such areas provide rich directions for future research, particularly at the intersection of participatory methods and social science research.

What specific concerns are raised during the deliberative process, and how are these addressed?

Diverse Voices panels[338] with experiential experts (both from non-profits and from within the community) may also be used to identify potential issues by asking questions such as "What mistakes could be made by decision makers because of how this proposal is currently worded?" and, "What does the proposal not say that you wish it said?" Other strategies we discussed for understanding multiple viewpoints may also have a role to play here when tailored towards sharing concerns. However, we also stress the importance of ongoing partnerships with impacted community members beyond just an initial session. By ensuring that the community has a voice throughout the project's lifetime, their values are always kept front-and-center. As others have argued, it is not sufficient to interview impacted communities once simply to "check the participatory box"[276].

8.6.1 Determining Capability Sets

Now, equipped with some guiding questions and concrete examples of a participatory approach, we will apply these principles directly to selecting capability sets at the outset of an AI4SG project. To identify what capabilities to work towards, some scholars such as Nussbaum[223] have proposed well-defined sets of capabilities and argued that society ought to focus on enabling those capabilities. However, we endorse the view that the selection of an optimal set of capabilities should be based on community consensus[264]. We argue that an AI project can only be for social good if it is responsive to the values of the communities affected by the AI system.

What functionings do members of the various stakeholder groups wish they could achieve through the implementation of the project?

In the fair machine learning literature, Martin Jr. et al.[197] propose a participatory method called community-based system dynamics (CBSD) to bring stakeholders in to help formulate a machine learning problem and mitigate bias. Their method is designed to understand causal relationships, specifically feedback loops, particularly those in high-stakes environments such as health care or criminal justice, that may disadvantage marginalized and vulnerable groups. This process is intended to bring in relevant stakeholders and recognize that their lived experience makes these participants more qualified to recognize the effects of these interventions. Using visual diagrams designed by impacted community members, the CBSD method can help identify levers that may help enable or inhibit functionings, particularly for those who are most vulnerable. Similarly, Simonsen & Robertson[273] suggest the use of mock-ups and prototypes to facilitate communication between experiential experts and developers. Other strategies discussed for understanding multiple viewpoints may also apply if tailored towards determining capabilities.

What functionings are the priority for those most vulnerable? Is there an overlap between their priorities and the goals of other stakeholders?

We need to pay attention to those who are most vulnerable. The capabilities approach may be leveraged to fight inequality by thinking in terms of capability equalizing. To identify capabilities that are not yet available to marginalized members of a community, we must listen to their concerns and ensure those concerns are prioritized, for example via strategies proposed in our discussions on finding those impacted by AI systems and including multiple viewpoints.

As an example of the consequences of failing to include those most vulnerable throughout the lifetime of an AI system, consider one of the initial steps of data collection. Women are often ignored in datasets and therefore their needs are underreported[86]. For example, crash-test dummies were designed to resemble the average male, and vehicles were evaluated to be safe based on these male dummies—leaving women 47% more likely to be seriously injured in a crash[236]. These imbalances are often also intersectional, as Buolamwini & Gebru[55] demonstrate by revealing stark racial and gender-based disparities in facial recognition algorithms.

Beyond the inclusion of all groups in datasets, data must be properly stratified to expose disparities. During the COVID-19 pandemic in December 2020, the Bureau of Labor Statistics reported a loss of 140,000 net jobs. The stratified data reveal that all losses were women's: women lost 156,000 jobs while men gained 16,000, and unemployment was most severe for Black, Latinx, and Asian women[91]. Without accounting for the capabilities of all people affected by such systems, it is difficult to claim that these technologies were for social good.

On the other hand, prioritizing the needs of the most marginalized groups may at times offer an accelerated path towards achieving collective goals. Project Drawdown identified and ranked 100 practical solutions for stopping climate change[123]. Number 6 on its list was the education of girls, recognizing that women with higher levels of education marry later and have fewer children, as well as manage agricultural plots with greater yield. Another solution advocates for securing land tenure of indigenous peoples, whose stewardship of the land fights deforestation, resource extraction, and monocropping.

Are any of these capability sets fertile, in the sense of securing other capabilities?

To maximize the capabilities of various communities, we may wish to focus on capabilities that produce fertile functionings, as discussed in Section 8.5. Specifically, many functionings are necessary inputs to produce others; for example, achieving physical fitness from playing sports requires as input good health and nourishment[69]. Some of Nussbaum's 10 central capabilities—including bodily integrity (security against violence and freedom of mobility) and control over one's environment (right of political participation, to hold property, and dignified work)—may be viewed as fertile[223]. Reading, for instance, may secure the capability to work, associate with others, and have control over one's environment. AI has the potential to help achieve many of these capabilities. For example, accessible crowdwork, done thoughtfully, offers the opportunity for people with disabilities to find flexible work without the need for transit[347].

How closely do the values of the project match those of the community as opposed to the designers? How does the focus of the project respond to their expressed needs and concerns? Does the project have the capacity to respond to those needs?

Consider the scenario where a community finds it acceptable to hunt elephants, while the designers are trying to prevent poaching. There could be agreement on high-level values, such as public health, but disagreement on whether to prioritize specific interventions to promote public health. There could even be a complete lack of interest in the proposed AI system by the community.

At an early stage of a project, AI researchers need to facilitate a consultation method to understand communities' values and choices. Note that this process could allow stark differences in priority between stakeholders to surface which prevent the project from starting so as not to over-invest in a project that will later be terminated because of difference in values. We may consider several of the strategies we discussed previously in this case, such as deliberative democratic processes, Diverse Voices panels, or computational social choice methods. It may subsequently be necessary to end the project if these strategies do not work.

8.6.2 Evaluating AI for Social Good

Once AI researchers and practitioners have a system tailored to these capabilities, we believe that communities should be the ones to judge the success of this new system. This stage of the process may pose additional challenges, given the difficulty of measuring capabilities[142]. Though we do not here endorse any measurement methodology, various attempts to operationalize the capabilities approach give us confidence that such methodologies are feasible and may be implemented in the course of evaluating AI projects[11].

First and foremost, we maintain that the evaluation of success should be done throughout the lifecycle of the AI4SG project (and beyond) as discussed above, especially via community feedback. However, we wish to emphasize that we as AI researchers need to keep capabilities in mind as we evaluate the success of AI4SG projects to avoid "unintended consequences" and a short-sighted focus on improved performance on metrics such as accuracy, precision, or recall.

How does the new AI4SG system affect all stakeholders' capabilities, particularly those selected at the start?

This question is related to the literature on measuring capabilities[142], and is therefore difficult to answer. We will aim to provide a few examples, which may not apply well to every AI4SG system, nor will it be an exhaustive list of techniques that could be valid. First, based on the idea of AI as a diagnostic to measure social problems[3], we may be able to (partially) probe this question using data. Sweeney[283] show through ad data that advertisements related to arrest records were more likely to be displayed when searching for "Black-sounding" names, which may affect a candidate's employment capability, for example. Obermeyer et al.[225] analyze predictions, model inputs and parameters, and true health outcome data to show that the capability of health is violated for Black communities, as they are included in certain health programs less frequently than white communities due to a biased algorithm.

There could additionally be feedback mechanisms when an intervention is deployed, whether via the AI system itself, or possibly by collaborating with local non-profits and NGOs. This may be especially useful in cases where these organizations are already engaged in tracking and improving key capabilities such as health outcomes, e.g., World Health Partners[62] or CARE international[139]. Again, these examples will likely not apply to all cases and opens the door for further, interdisciplinary research. However, no matter what strategy is taken, it is imperative that we continue to center communities in all attempts to measure the effects on stakeholders' capabilities.

Are other valued capabilities or functionings negatively affected as a result of the project? Are stakeholders' values and priority rankings in line with such tradeoffs?

We have a responsibility to *actively* think about possible outcomes; it is neglect to dismiss negative possibilities as "unintended consequences"[234]. These participatory mechanisms should thus ensure that the perspective of the most vulnerable and most impacted stakeholders is given due consideration. We recognize that this can be especially challenging, as discussed in our first guiding question for identifying impacted communities. Therefore, we further suggest that the evaluation of an AI4SG project should employ consultation mechanisms that are open to all community members throughout the implementation process, such as the feedback mechanisms suggested previously.

What should my role be as an AI researcher? As a student?

We believe that AI researchers at all levels should participate in this work. This work involves all of the above points, including learning about the domain to understand who stakeholders are, discussing with the stakeholders, and evaluating performance. We also acknowledge that we AI researchers may not always be the best suited to lead participatory efforts, and so encourage interdisciplinary collaborations between computer science and other disciplines, such as social sciences. A strong example would be the Center for Analytical Approaches to Social Innovation (CAASI) at the University of Pittsburgh, which brings together interdisciplinary teams from policy, computing, and social work[56]. However, AI researchers should not completely offload these important scenarios to social science or non-profit colleagues. It should be a team effort, which we believe will bring fruitful research in social science, computer science, and even more disciplines.

AI researchers and students can also advocate for systematic change from within, which we discuss more in depth in the conclusion. Although student researchers are limited by constraints such as research opportunities and funding, they may establish a set of moral aspirations for their work and set aside time for people-centered activities, such as mentoring and community-building[61].

8.7 Conclusion: Thoughts on AI for Social Good as a Field

In this paper, we lay out a community-centered approach to defining AI for social good research that focuses on elevating the capabilities of those members who are most marginalized. This focus on capabilities, we argue, is best enacted through a participatory approach that includes those affected throughout the design, development and deployment process, and gives them ground to choose their desired capability sets as well as influence how they wish to see those capabilities realized.

We recognize that the participatory approach we lay out requires a significant investment of time, energy, and resources beyond what is typical in AI or even much existing AI4SG research. *We high-*

light this discrepancy to urge a reformation within the AI research community to reconsider existing incentives to encourage researchers to pursue more socially impactful work.

Institutions have the power to catalyze change by (1) establishing requirements for community engagement in research related to public-facing AI systems; and (2) increasing incentives for researchers to meaningfully engage impacted communities while simultaneously producing more impactful research[32].

While engaging in collaborative work with communities can give rise to some technical directions of independent interest to the AI community[81], such a shift to encourage community-focused work will in part require reconsidering evaluation criteria used when reviewing papers at top AI conferences. Greater value must be placed on papers with positive social outcomes, including those with potential for impact, if the work has not yet been deployed. Such new criteria are necessary since long-term, successful AI4SG partnerships often also lead to non-technical contributions, as well as situated programs which do not focus necessarily on generalizability[237]. We encourage conferences to additionally consider rewarding socially beneficial work with analogies to Best Paper awards, such as the "New Horizons" award from the MD4SG 2020 Workshop, and institutions to recognize impactful work such as the Social Impact Award at the Berkeley School of Information[215].

In the meantime, we suggest that researchers look to nontraditional or interdisciplinary venues for publishing their impactful community-focused work. These venues often gather researchers from a variety of disciplines outside computer science, opening the door for future collaborations. For example, researchers could consider COMPASS, IAAI, MD4SG/EAAMO[4], the LIMITS workshop[212], and special tracks at AAAI and IJCAI, among others. Venues such as the Computational Sustainability Doctoral Consortium and CRCS Rising Stars workshop bring students together from multiple disciplines to build relationships with each other. Researchers could also consider domain workshops and conferences, such as those in ecology or public health.

The incentive structure in AI research is often stacked against thoughtful deployment. Whereas a traditional experimental section may take as little as a week to prepare, a deployment in the field may take months or years, but is rarely afforded corresponding weight by reviewers and committee members. This extended timeline weighs most heavily on PhD students and untenured faculty who are evaluated on shorter timescales. We should thus reward both incremental and long-term deployment, freeing researchers from the pressure to rush to deployment before an approach is validated and ready.

In addition to the need for bringing stakeholders into the design process of AI research, we must ensure that all communities are welcomed as AI researchers as well. Such an effort could counteract existing disparities and inequities within the field. For example, as is the case in other academic disciplines, systemic anti-Blackness is ingrained in the AI community, with racial discrepancies in physical resources such as access to a secure environment in which to focus, social resources such as access to project collaborators or referrals for internships, and measures such as the GRE or teacher evaluations[118]. Further, as of 2018, only around 20% of tenure-track computer science faculty were women[261]. To combat these inequities, people across the academic ladder may actively work to change who they hire, with whom they collaborate, including collaborations with minority-serving institutions[166], and how much time they spend on service activities to improve diversity efforts[118].

The above reformations could contribute greatly to making the use of participatory approaches the norm in AI4SG research, rather than the exception. PACT, we argue, is a meaningful new way to answer the question: “what is AI for social good?” We, as AI researchers dedicated to the advancement of social good, must make a PACT with communities to find our way forward together.

9

Conclusion

To conclude, I believe that it is very important to take a holistic view of the AI for social impact pipeline, from gathering and analyzing data, to multi-agent reasoning, to deployment. The holistic view allows us to identify areas with uncertainty, and plan to account for that in other portions of the pipeline. This requires us to use multiple methodologies throughout. I have given several examples, including machine learning, domain adaptation, and game theory, to name a few. The holistic view also allows us to acknowledge the role of humans in these systems, which requires us to not only reason about human behavior, including through human subject experiments, but also to be more inclusive as we design and deploy these systems.

My vision going forward is to use this holistic view of AI for social impact systems to support local communities, including domain and experiential experts and those impacted by systems, in addressing difficult challenges, particularly in areas like conservation, public health, and education. Technically, I believe AI for social impact is dependent on systems with efficient data analysis and strategic reasoning, while accounting for humans-in-the-loop and other real-world challenges such as uncertainty. There are several open questions I will pursue to achieve these goals:

Multi-modal and limited data: It can take a vast amount of human effort to collect real-world data, yet there are ever-increasing amounts of data in a multitude of modalities. I strive to lessen the amount of human effort required to curate datasets, an endeavor I have begun with VIOLA[39], by leveraging multiple modalities and data augmentation, as I have started with thermal data[37,38,40] and MND prediction with satellite data[36]. In an ongoing project, I am building an active learning methodology to strategically collect labeled data while incorporating important real-world criteria such as fairness, and multiple imaging modalities such as thermal, RGB, and Lidar. Further work on semi-supervised learning and theoretical work on strategically collecting data via multi-armed bandits and reinforcement learning, like my prior work on allocation[42,303,328], is also important.

Human-AI collaboration: Given the important, often safety-critical role that AI systems for social impact may assume, humans remain a vital part of deployed systems. Humans may play different roles in such systems, for example providing inputs, being affected by outputs, making decisions from recommendations, or relying on AI recommendations until some uncertainty arises, as in the case of selective prediction. Existing work in AI largely ignores these varied forms of interaction, and often makes simplifying assumptions when modeling human behavior, such as rationality assumptions. As such, there remain a plethora of questions in how humans and AI systems can collaborate, especially to complement one another. I aim to expand my work in selective prediction[41] to model human-AI interactions over time, e.g., as humans grow tired or accustomed to the system, and to consider bounded rationality in new human-AI systems.

Deployment: In my immediate next steps, I aim to enact the PACT framework[43]. This will require building further community partnerships, and conducting research to best address the practical challenges in using PACT, such as finding consensus among stakeholders, supporting expeditious issue reporting, and respecting stakeholders' autonomy and privacy in AI systems. I believe that this also requires contributing to an education system where everyone, including those from historically underrepresented groups, can pursue their interests, and will also continue my efforts with Try AI as a result.

Overall, I believe that AI has a great deal of potential for positive impact in the world, but I believe that we must work together with people from many fields and backgrounds, including domain experts, students, and those impacted by AI systems, to ensure that AI is truly an opportunity for humanity.

A

Appendix to Chapter 1

A.1 EFG and POMDP

We first discuss in more detail how we exploit the structure of our game and provide a scalable algorithm by extending the use of coverage probabilities and multiple LPs, instead of using an EFG. The game tree is shown in Fig. A.1. Our approach solves at most $8N$ LPs with $O(N + |\mathcal{E}|)$ constraints and $O(N + |\mathcal{E}|)$ variables, where $\mathcal{E}$ is the defender pure strategy set. Using the EFG approach, the size of the game tree is $O(N \cdot 4^l \cdot |\mathcal{E}|)$, where l is the number of drones. The EFG multiple LPs approach therefore solves exponentially more LPs, each with a much larger size than ours. One might also consider using a POMDP to model the movement of the defender from allocation to a new reaction target with the unobservable state being whether an adversary is present or not. However, a POMDP model does not capture all of the intricacies due to strategic game interactions. For example, it does not account for the fact that the adversary will choose a location to attack rationally.

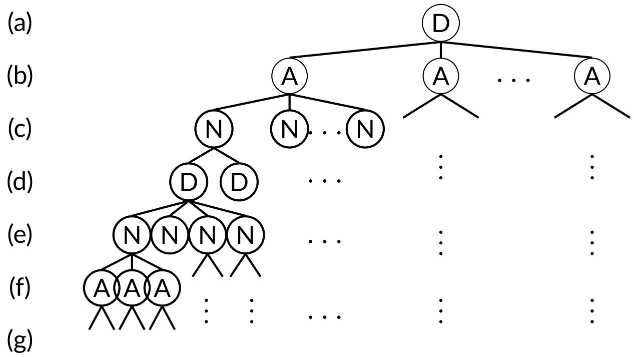


Figure A.1: Game tree illustrating defenders' (D) allocation, signaling, and reaction steps, as well as those of the adversary (A) and nature (N) (uncertainty). Level (a) is the initial defender deployment, (b) is the adversary choice of target, (c) is the detection (with uncertainty), (d) is the defender signaling, (e) is the adversary's observation (with uncertainty), (f) is the adversary's decision to run away or not, (g) is when players receive payoffs.

A.2 Omitted Proofs in Section 1.4

Proposition 1. *Let* $\chi_0^* = \chi^*(0)$ *be the defender optimal deployment when no uncertainties exist. There exist instances where* $\mathsf{DefEU}(\chi_0^*, \gamma) < \mathsf{DefEU}(\chi^*(\gamma), \gamma)$ *for some* γ.

Proof. We prove by constructing such an instance. Consider the graph in Fig. A.2 with 4 targets, 1 human patroller, and 2 sensors. The adversary chooses one target to attack. A successful attack gives the adversary utility of $+2$ and defender utility of -5, whereas catching the adversary yields adversary utility -1 and defender utility 0. If the adversary chooses not to attack after observing a signal from the sensor, both the adversary and the defender receive 0.

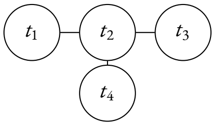


Figure A.2: Diagram of detection uncertainty example.

If sensors have perfect detection ($\gamma = 0$), we can place the human patroller at t_2, place the two sensors at t_1 and t_3 respectively, and match the patroller to t_4. Thus, we cover all targets with probability 1. When we observe the adversary, we will always send the strong signal (σ_1) to ensure the adversary will run away. Therefore, the adversary is better off not attacking, yielding utility of 0 for both.

To see how a false negative detection can affect the solution quality, we now consider the case with $\gamma = 0.5$. If we use the same strategy but with imperfect detection, the adversary can attack t_1 or t_3 successfully with probability $1/2$ (when the adversary does not observe a warning signal σ_1), and run away with probability $1/2$ (when the adversary observes the signal σ_1). The defender's expected utility is $\mathsf{DefEU}(\cdot) = \frac{1}{2}(0) + \frac{1}{2}(-5) = -\frac{5}{2} = -2.5$.

We want to show the optimal strategy when $\gamma = 0.5$. First, we will always have a patroller at t_2. We will always have one drone at t_1 and we will have another drone at t_3 with probability $6/9$ and at t_4 with probability $3/9$. We will match the patroller to t_1 with probability $2/9$; t_3 with probability $3/9$ ($1/9$ when there's sensor, $2/9$ when there's nothing); t_4 with probability 4/9 when there's nothing. We will first calculate defender expected utility at t_1, t_2 and t_4. At target t_1, the adversary always observe a sensor, with probability $7/9$ it's state s$-$ and with probability $2/9$ it's state s$+$. Thus $\mathsf{AttEU}(t_1) = 2/9 \cdot (-1) + 7/9 \cdot (1-\gamma)(-1) + 7/9 \cdot \gamma \cdot (2) = 1/6$, thus the adversary will attack. For target t_2 the adversary will always get caught, so he won't attack. For target t_4, if he observes nothing, he will get caught with probability $\frac{4/9}{4/9+2/9} = 2/3$, so he will not attack when he observes nothing. If he observes a drone, he will only get caught with probability $(1-\gamma)$, thus gains utility of $\gamma \cdot 1/3 = 1/6$.

Consider the following signaling scheme at t_3. If we detect an adversary or the state is matched, then we will send σ_0, if we don't detect and state is not matched we will send signal with marginal probability $\phi^{\mathrm{s}-} = 7/18$. If the adversary does not observe a drone, he will not attack. If he observes a drone with σ_0, then the adversary expected utility is $((1-\gamma)\cdot 2/3)\cdot(-1)+\gamma\cdot 1/9\cdot(-1)+\gamma\cdot 7/18\cdot(2) = 0$. Thus the adversary only attacks when he observes σ_1 this happens with probability $1/12$, thus the adversary gets expected utility of $2\cdot 1/12 = 1/6$. Since we are getting attacked with probability $1/12$, the defender expected utility is now $-5\cdot 1/12 = -0.416$ thus doing getting better expected utility by considering uncertainty.

With optimal deployment, we can get a defender expected utility of -0.416 when $\gamma = 0.5$. This example shows that when the detection uncertainty does not exist, a very simple deployment yields the optimal expected utility. However, this strategy is no longer optimal when detection uncertainty is present. Therefore, we need to consider detection uncertainty and compute the new optimal solution.

□

Theorem 1. $\mathsf{DefEU}(\chi^*(\gamma),\gamma) \geq \mathsf{DefEU}(\chi^*(\gamma'),\gamma')$ *for any* $\gamma' > \gamma$ *in any problem instance.*

Proof. Let $\chi^*_\gamma = \chi^*(\gamma)$. Throughout the proof, we assume no observational uncertainty. Assume for the contradiction that defender expected utility strictly increases as detection uncertainty increases, i.e., $\mathsf{DefEU}(\chi^*_\gamma,\gamma) < \mathsf{DefEU}(\chi^*_{\gamma+\varepsilon},\gamma+\varepsilon)$ for some ε. Let ψ and ϕ be the corresponding signaling variables from $\chi^*_{\gamma+\varepsilon}$.

Consider the following new variables χ' and $\psi' = \frac{(1-\gamma-\varepsilon)\psi+\varepsilon\phi}{1-\gamma}$, and let all other variables be the same as $\chi^*_{\gamma+\varepsilon}$. First, note that $\psi'^{\theta} = \frac{(1-\gamma-\varepsilon)\psi^{\theta}+\varepsilon\phi^{\theta}}{1-\gamma} \leq \frac{(1-\gamma-\varepsilon)x^{\theta}+\varepsilon x^{\theta}}{1-\gamma} = x^{\theta}$, for all $\theta \in \{\mathrm{s}+,\mathrm{s}-,\bar{\mathrm{s}}\}$; therefore, all variables are feasible. Thus, we have $\mathsf{DefEU}(\chi',\gamma) \leq \mathsf{DefEU}(\chi^*_\gamma,\gamma) < \mathsf{DefEU}(\chi^*_{\gamma+\varepsilon},\gamma+\varepsilon)$. Furthermore, consider an augmented strategy where when we observe an adversary at state $\psi'^{\mathrm{s}-}$, with marginal probability $\phi^{\mathrm{s}-}$, we ignore the detection, thus make the target uncovered. Note that this strategy is still feasible, and makes our defender expected utility lower.

Now, we will calculate the defender expected utility when the defender allocates security resources according to χ'. We can decompose the expected utility by different signals, i.e. $\mathsf{DefEU}(\chi', \gamma) = \sum_{\omega \in \{\mathrm{n}, \sigma_0, \sigma_1\}} \mathsf{DefEU}(\chi', \gamma | \omega)$, where $\mathsf{DefEU}(\chi', \gamma|\omega)$ is the defender expected utility given state ω. First, note that $\mathsf{DefEU}(\chi', \gamma|\mathrm{n})$ stays the same as detection uncertainty changes. Thus, we will only look at $\mathsf{DefEU}(\chi', \gamma|\sigma_0)$ and $\mathsf{DefEU}(\chi', \gamma|\sigma_1)$.

$$
\begin{aligned}
&\mathsf{DefEU}(\chi', \gamma|\sigma_0)\\
&\qquad = (1-\gamma) \cdot (\psi'^{\mathrm{s}+} U^d_+ + \psi'^{\mathrm{s}-} U^d_+ + \psi'^{\bar{\mathrm{s}}} U^d_-)\\
&\qquad\quad + \gamma \cdot (\phi^{\mathrm{s}+} U^d_+ + \phi^{\mathrm{s}-} U^d_- + \phi^{\bar{\mathrm{s}}} U^d_-)\\
&\qquad \geq (1-\gamma-\varepsilon) \cdot (\psi^{\mathrm{s}+} U^d_+ + \psi^{\mathrm{s}-} U^d_+ + \psi^{\bar{\mathrm{s}}} U^d_-)\\
&\qquad\quad + \varepsilon(\phi^{\mathrm{s}+} U^d_+ + \phi^{\mathrm{s}-} U^d_- + \phi^{\bar{\mathrm{s}}} U^d_-) \qquad\qquad \text{(by our augmented strategy)}\\
&\qquad\quad + \gamma \cdot (\phi^{\mathrm{s}+} U^d_+ + \phi^{\mathrm{s}-} U^d_- + \phi^{\bar{\mathrm{s}}} U^d_-)\\
&\qquad = (1-\gamma-\varepsilon) \cdot (\psi^{\mathrm{s}+} U^d_+ + \psi^{\mathrm{s}-} U^d_+ + \psi^{\bar{\mathrm{s}}} U^d_-)\\
&\qquad\quad + (\gamma+\varepsilon) \cdot (\phi^{\mathrm{s}+} U^d_+ + \phi^{\mathrm{s}-} U^d_- + \phi^{\bar{\mathrm{s}}} U^d_-)\\
&\qquad = \mathsf{DefEU}(\chi^*_{\gamma+\varepsilon}, \gamma+\varepsilon|\sigma_0)
\end{aligned}
$$

We also want to show that $\mathsf{DefEU}(\chi', \gamma|\sigma_1) = \mathsf{DefEU}(\chi^*_{\gamma+\varepsilon}, \gamma+\varepsilon|\sigma_1)$. Recall that they are both 0 because the adversary will run away. Then, we have $\mathsf{DefEU}(\chi', \gamma) \geq \mathsf{DefEU}(\chi^*_{\gamma+\varepsilon}, \gamma+\varepsilon) > \mathsf{DefEU}(\chi^*_\gamma, \gamma)$. This contradicts χ^*_γ is an optimal solution.

□

Proposition 2. *$\chi^*(\gamma)$ differs from $\chi^*(\gamma')$ for any $\gamma' > \gamma$ when x^{s-}_t is nonzero for $\chi^*(\gamma')$, where target*

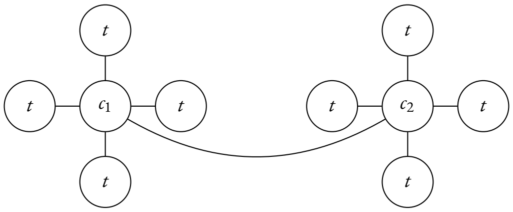
t is the adversary best responding target in $\chi^*(\gamma')$.

Proof. Suppose for contradiction that χ is an optimal solution for both $\mathsf{DefEU}(\chi,\gamma)$ and $\mathsf{DefEU}(\chi,\gamma+\varepsilon)$. Consider χ' that is obtained same way as previous proof. i.e. $\psi' = \frac{(1-\gamma-\varepsilon)\psi+\varepsilon\phi}{1-\gamma}$ and all other variables stays the same.

Note $\mathsf{DefEU}(\chi',\gamma)$ is strictly bigger than $\mathsf{DefEU}(\chi,\gamma+\varepsilon)$ if $\psi^{s-}, x^{s-}-\psi^{s-}, \phi^{s-}$ or $x^{s-}-\phi^{s-}$ is non-zero. In other words, if x^{s-} is nonzero, the $\mathsf{DefEU}(\chi',\gamma) > \mathsf{DefEU}(\chi,\gamma+\varepsilon)$, thus contradicts $\mathsf{DefEU}(\chi,\gamma)$ is an optimal solution. □

Proposition 3. *There exists Π such that the loss due to ignoring observational uncertainty is arbitrarily large. In other words,* $\mathsf{DefEU}(\chi^*(\gamma_0,\Pi),\gamma_0,\Pi)$ - $\mathsf{DefEU}(\chi^*(\gamma_0,\Pi_0),\gamma_0,\Pi) > M$*,* $\forall M > 0$.

Proof. We will show an example where the new signaling strategy is arbitrarily better than the naive signaling strategy. Consider the following example. We have 10 targets, 1 human patroller, and 8 sensors.



The optimal allocation strategy is to allocate sensors in all t's, allocate the human patroller in one of the center vertices (c_1, c_2) uniformly at random, and match to the other center vertex. For all targets, utility is defined as the following, for some arbitrarily big M. Note that the adversary's expected utilities for attacking c_1 and c_2 are both 0.

	covered	uncovered
Adversary	-1	$1+\varepsilon$
Defender	0	$-M$

Consider the following uncertainty matrix. Let $r_0 = \frac{1-\varepsilon}{1+\varepsilon} \cdot (1-2\varepsilon')$, for some $\varepsilon' > 0$.

$\Pr[\hat{\omega}\vert\omega]$	$\omega = \mathrm{n}$	$\omega = \sigma_0$	$\omega = \sigma_1$
$\hat{\omega} = \mathrm{n}$	1	$1 - r_0$	ε'
$\hat{\omega} = \sigma_0$	0	r_0	$1 - 2\varepsilon'$
$\hat{\omega} = \sigma_1$	0	0	ε'

Let r_{p} be the probability of state p, let r_{n} be the probability of state n+ and n−, and let $r_{\mathrm{s}+}$ and $r_{\bar{\mathrm{s}}}$ be the probability of the state s+ and s− and $\bar{\mathrm{s}}$, respectively. Let $r_o = \frac{1}{1+\varepsilon}$ (the optimal signaling strategy ignoring uncertainty). In this strategy, the adversary attacks when he observes state n and σ_0. Therefore, $\bar{\eta} = [1, 1, 0]$ is the vector that depicts adversary behavior in this case.

$$
\begin{aligned}
&\mathsf{DefEU}(r_o)\\
&\quad = U^d_+ r_{\mathrm{p}} + U^d_- r_{\mathrm{n}} + U^d_- r_{\bar{\mathrm{s}}}(1 - r_n)\bar{\eta} \cdot \Pr[\hat{\omega}|\omega = \sigma_0]\\
&\quad + U^d_- r_{\bar{\mathrm{s}}}(r_n)\bar{\eta} \cdot \Pr[\hat{\omega}|\omega = \sigma_1] + U^d_+ r_{\mathrm{s}+}\bar{\eta} \cdot \Pr[\hat{\omega}|\omega = \sigma_1]\\
&\quad = -\frac{1}{2} - \frac{M-1}{2}(1-\varepsilon') \leq -\frac{M-2}{2}(1-\varepsilon')
\end{aligned}
$$

Let $r_n = \varepsilon'$ be the new signaling strategy. Let η' be the new adversary's attacking vector. Observe the adversary will only attack when he observes state n. Therefore, $\eta' = [1, 0, 0]$.

$$
\begin{aligned}
&\mathsf{DefEU}(r_n)\\
&\quad = U^d_+ r_{\mathrm{p}} + U^d_- r_{\mathrm{n}} + U^d_- r_{\bar{\mathrm{s}}}(1 - r_n)\eta' \cdot \Pr[\hat{\omega}|\omega = \sigma_0]\\
&\quad + U^d_- r_{\bar{\mathrm{s}}}(r_n)\eta' \cdot \Pr[\hat{\omega}|\omega = \sigma_1] + U^d_+ r_{\mathrm{s}+}\eta' \cdot \Pr[\hat{\omega}|\omega = \sigma_1]\\
&\quad = -\frac{M}{2}(1-\varepsilon')(1 - \frac{M-2}{M}(1-2\varepsilon')) - \frac{M}{2}\varepsilon'\varepsilon'\\
&\quad \geq -(1-\varepsilon') - \frac{M}{2}\varepsilon'\varepsilon'
\end{aligned}
$$

For $0 < \varepsilon' < 1/M$, we get the gap of $O(M)$.

□

Theorem 2. *For any fixed deployment χ, if the adversary's best response is $(t, 0)$ or $(t, 1)$ at the Stackelberg equilibrium with Π_0, then it stays as an equilibrium for any Π'.*

Proof. The proof of Theorem 2 follows from this Lemma:

Lemma 1. *For any fixed χ, if $(t, 1)$ (or $(t, 0)$) is a best response for the adversary at Π_0, then $(t, 1)$ (or $(t, 0)$) is also a best response for all Π', for any $\Pi' \neq \Pi_0$.*

Proof. The proof of the Lemma follows from the following two claims. Let $\mathsf{DefEU}(\chi, t, \eta, \Pi)$ be the defender's expected utility when she plays the deployment χ. We add t and η to the typical notation to represent that the adversary's strategy is to attack t with behavior η, and the observational uncertainty matrix is Π. There is no detection uncertainty. We use a similar notation for $\mathsf{AttEU}(\chi, t, \eta, \Pi)$. Let us also index η with $\hat{\omega} \in \Omega$ as $\eta^{\hat{\omega}}$, and reference η for a target, i, as η_i, for a final notation of $\eta_i^{\hat{\omega}}$. For example, to reference the adversary behavior for $\hat{\omega} = \mathrm{n}$ and target i, we can write it as η_i^{n}.

Claim 1. *For any χ, t, η, $\mathsf{AttEU}(\chi, t, \eta, \Pi') \leq \mathsf{AttEU}(\chi, t, \eta^*_{0,t}, \Pi_0)$, where $\eta^*_{0,t}$ is the best response when $\Pi = \Pi_0$.*

Proof. Let $\mathsf{AttEU}(\omega) = \Pr[\omega]\mathsf{AttEU}(\chi, t, 1)$ be the adversary's expected utility when true signaling state is ω and the adversary attacks the target t.

$$\begin{aligned}
&\mathsf{AttEU}(\chi, t, \eta, \Pi') = \\
&\qquad \sum_{\hat{\omega}\in\Omega} \eta^{\hat{\omega}} \cdot \left(\sum_{\omega\in\Omega} \Pr[\hat{\omega}|\omega] \mathsf{AttEU}(\omega) \right) \\
&\qquad = \sum_{\omega\in\Omega} \mathsf{AttEU}(\omega) \left(\sum_{\hat{\omega}\in\Omega} \eta^{\hat{\omega}} \Pr[\hat{\omega}|\omega] \right) \\
&\qquad \leq \quad \sum_{\omega\in\Omega} \mathsf{AttEU}(\omega) \cdot 1\left(\mathsf{AttEU}(\omega) \geq 0\right) \\
&\qquad \leq \quad \mathsf{AttEU}(\chi, t, \eta^*_{0,t}, \Pi_0) \qquad\qquad (\text{Note } \Pi_0 \text{ means } \Pr[\hat{\omega}|\omega] = 1 \text{ for all } \hat{\omega} = \omega)
\end{aligned}$$

□

Where $1(\cdot)$ is an indicator function, $1(\cdot) = 1$ if the corresponding expression is true, and $1(\cdot) = 0$ otherwise. Note that $\mathsf{AttEU}(\chi, t, \eta, \Pi') \leq \mathsf{AttEU}(\chi, t, \eta, \Pi_0)$ is not true. Consider a Π_i which is some permutation matrix of $\mathbf{I}$.

Claim 2. *For any χ and t, the adversary's expected adversary utility stays the same for any Π if the adversary behavior is 1 or 0. In other words, $\mathsf{AttEU}(\chi, t, 1, \Pi') = \mathsf{AttEU}(\chi, t, 1, \Pi_0)$ and $\mathsf{AttEU}(\chi, t, 0, \Pi') = \mathsf{AttEU}(\chi, t, 0, \Pi_0)$.*

Corollary 1. *We also have $\mathsf{DefEU}(\chi, t, 1, \Pi') = \mathsf{DefEU}(\chi, t, 1, \Pi_0)$ and $\mathsf{DefEU}(\chi, t, 0, \Pi') = \mathsf{DefEU}(\chi, t, 0, \Pi_0)$.*

Proof. If $\eta = 1$ then $\sum_{\hat{\omega}\in\Omega} \eta^{\hat{\omega}} \Pr[\hat{\omega}|\omega] = 1$ for all $\omega \in \Omega$ independent of Π. Therefore, we get $\mathsf{AttEU}(\chi, t, a, \Pi) = \sum_{\omega\in\Omega} \mathsf{AttEU}(\omega)$ independent of Π, and the claim holds.

Similarly, if $\eta = 0$ then $\sum_{\hat{\omega}\in\Omega} \eta^{\hat{\omega}} \Pr[\hat{\omega}|\omega] = 0$ for all $\omega \in \Omega$ independent of Π.

Exactly the same argument holds for calculating $\mathsf{DefEU}(\cdot)$. □

By combining the two claims we get the following: $\mathsf{AttEU}(\chi, t, \eta^*_{0,t}, \Pi_0) = \mathsf{AttEU}(\chi, t, 1, \Pi_0) = \mathsf{AttEU}(\chi, t, 1, \Pi') \geq \mathsf{AttEU}(\chi, t, \eta, \Pi')$, thus we get $(t, 1)$ as a best response for Π', for any Π' and χ. □

This shows if $(t, 1)$ or $(t, 0)$ is a Stackelberg equilibrium, the defender can safely deploy the same strategy for any uncertainty matrix Π', without any loss in her expected utility.

□

Theorem 3. *If $(t, 1)$ is a best response for $\Pi_{\kappa\lambda\mu}$ and χ is a weak-signal-attack deployment, then $(t, 1)$ is a best response for $\Pi_{\kappa'\lambda'\mu'}$ and χ for all $\kappa' \geq \kappa$, $\lambda' \geq \lambda$, $\mu' \geq \mu$.*

Let $\mathsf{AttEU}(\hat{\omega}) = \Pr[\hat{\omega}]\mathsf{AttEU}(\chi, t, 1)$ be the adversary's expected utility when observed signaling state is $\hat{\omega}$ and the adversary attacks the target t.

Proof. We have $\eta^{\sigma_1} = 1$, which implies $\mathsf{AttEU}(\hat{\omega} = \sigma_1) \geq 0$ because of our Π structure. Therefore, increasing λ or μ only increases $\mathsf{AttEU}(\hat{\omega} = \mathrm{n})$ and $\mathsf{AttEU}(\hat{\omega} = \sigma_0)$, respectively. This implies $\mathsf{AttEU}(\hat{\omega} = \mathrm{n})$ and $\mathsf{AttEU}(\hat{\omega} = \sigma_0)$ stays positive. Therefore, the adversary behavior also stays the same, when we increase λ or μ.

Since χ is a weak-signal-attack deployment, we know $\mathsf{AttEU}(\hat{\omega} = \sigma_0) \geq 0$. Therefore, increasing κ only makes $\mathsf{AttEU}(\hat{\omega} = \mathrm{n})$, and $\mathsf{AttEU}(\hat{\omega} = \sigma_0)$ more positive; therefore, the adversary behavior stays as 1. □

Proposition 4. *There always exists an optimal solution that is a weak-signal-attack deployment with Π_0.*

Proof. Suppose for contradiction there does not exist an optimal solution that is a weak-signal-attack deployment. Then, consider the optimal solution χ^* with the least number of non-weak-signal-attack targets. By the assumption, we know $\mathsf{AttEU}(\sigma_0) < 0$ for some target t. Fix an arbitrary target t that is a non-weak-signal-attack target.

Then, we know $\mathsf{AttEU}(\sigma_0) = U^a_+(t) \cdot (\psi^{s+} + \psi^{s-}) + U^a_-(t) \cdot (\psi^{\bar{s}}) < 0$. Since $\mathsf{AttEU}(\sigma_0) < 0$ and $\Pi = \Pi_0$, we know the adversary is not attacking when he observes σ_0. Also, since $\mathsf{AttEU}(\cdot)$ is strictly negative, we know ψ^{s+} or ψ^{s-} is strictly greater than 0.

Consider the new deployment χ where we can decrease ψ^{s+} and/or ψ^{s-} until $\mathsf{AttEU}(\sigma_0) = 0$. Since the adversary is still not attacking when he observes σ_0 (recall we break ties in favor of the defender), the defender expected utility stays the same. Furthermore, this change only increases $x^{s+} - \psi^{s+}$ and $x^{s-} - \psi^{s-}$. Thus, $\mathsf{DefEU}(\sigma_1)$ also only increases. Our new χ is therefore still an optimal solution and t is now a weak-signal-attack target. Thus, it contradicts the assumption that χ^* is the optimal solution with the least number of non-weak-signal-attack targets. □

A.2.1 Handling Observational Uncertainty

The problem uses a similar linear program as for the case without observational uncertainty:

$$
\begin{aligned}
\max_{x,\psi,\phi} \quad & U^d_{\text{-}s}(t) + U^d_{\sigma_1}(t) + U^d_{\sigma_0}(t) && \text{(A.1)}\\
\text{s.t.} \quad & \textstyle\sum_{\mathbf{e}\in\mathcal{E}: e_i=\theta} q_{\mathbf{e}} = x_i^\theta \qquad \forall\,\theta \in \Theta, \forall\, i \in [N] && \text{(A.2)}\\
& \textstyle\sum_{\mathbf{e}\in\mathcal{E}} q_{\mathbf{e}} = 1 && \text{(A.3)}\\
& q_{\mathbf{e}} \geq 0 \qquad\qquad\qquad\qquad \forall\, \mathbf{e} \in \mathcal{E} && \text{(A.4)}\\
& U^a_{\hat{\omega}}(\psi_i, \phi_i, x_i) \leq b_i^{\hat{\omega}} \qquad\qquad \forall \hat{\omega} \in \Omega, \forall i \neq t && \text{(A.5)}\\
& 0 \leq b_i^{\text{n}} \quad 0 \leq b_i^{\sigma_0} \quad 0 \leq b_i^{\sigma_1} \qquad\qquad \forall i \neq t && \text{(A.6)}\\
& U^a_{\text{-}s}(t) + U^a_{\sigma_1}(t) + U^a_{\sigma_0}(t) && \\
& \quad \geq x_i^{\text{p}} \cdot U^a_{+}(i) + b_i^{\text{n}} + b_i^{\sigma_0} + b_i^{\sigma_1} \qquad \forall\, i \neq t && \text{(A.7)}\\
& 0 \leq \psi_i^\theta \leq x_i^\theta \quad \forall\, \theta \in \{\bar{\text{s}}, \text{s}-, \text{s}+\},\, \forall\, i \in [N] && \text{(A.8)}\\
& 0 \leq \phi_i^\theta \leq x_i^\theta \quad \forall\, \theta \in \{\bar{\text{s}}, \text{s}-, \text{s}+\},\, \forall\, i \in [N] && \text{(A.9)}\\
& (2\eta_t^{\hat{\omega}} - 1) \cdot U^a_{\hat{\omega}}(\psi_t, \phi_t, x_t) \geq 0 \qquad \forall\, \hat{\omega} \in \Omega && \text{(A.10)}
\end{aligned}
$$

However, here the utility functions need to be redefined in order to take observational uncertainty into account. We define $p^a_\omega(i) = \sum_{\hat{\omega}\in\Omega} \eta_i^{\hat{\omega}} \cdot \Pr[\hat{\omega}|\omega]$ as the probability of the adversary attacking target i given the true signaling state is $\omega \in \Omega$.

1. $U^{d/a}_{\text{-}s}(i) = x_i^{\text{p}} \cdot U^{d/a}_{+}(i) + x_i^{\text{n}+} \cdot U^{d/a}_{+}(i) \cdot \eta_i^{\text{n}} + x_i^{\text{n}-} \cdot U^{d/a}_{-}(i) \cdot \eta_i^{\text{n}}$ is the expected defender/adversary utility of target i being attacked over states when i has no sensor (p, n+, n−). This is nearly the same as the version with only detection uncertainty, but we must include the η since an adversary may run away when there is nothing when we consider observational uncertainty.

2. $U^{d/a}_{\sigma_0}(i) = (1-\gamma) \cdot p^a_{\sigma_0}(i) \cdot [\psi_i^{\text{s}+} \cdot U^{d/a}_{+}(i) + \psi_i^{\text{s}-} \cdot U^{d/a}_{+}(i) + \psi_i^{\bar{\text{s}}} \cdot U^{d/a}_{-}(i)] + \gamma \cdot p^a_{\sigma_0}(i) \cdot [\phi_i^{\text{s}+} \cdot U^{d/a}_{+}(i) + \phi_i^{\text{s}-} \cdot U^{d/a}_{-}(i) + \phi_i^{\bar{\text{s}}} \cdot U^{d/a}_{-}(i)]$ is the defender/adversary expected utility when the

adversary attacks target i and the defender signals σ_0. This has the added $p^a_{\sigma_0}(i)$ compared to the version with only detection uncertainty.

3. $U^{d/a}_{\sigma_1}(i) = (1-\gamma) \cdot p^a_{\sigma_1}(i) \cdot [(x^{\mathrm{s}+}_i - \psi^{\mathrm{s}+}_i) \cdot U^{d/a}_{+}(i) + (x^{\mathrm{s}-}_i - \psi^{\mathrm{s}-}_i) \cdot U^{d/a}_{+}(i) + (x^{\bar{\mathrm{s}}}_i - \psi^{\bar{\mathrm{s}}}_i) \cdot U^{d/a}_{-}(i)] + \gamma \cdot p^a_{\sigma_1}(i) \cdot [(x^{\mathrm{s}+}_i - \phi^{\mathrm{s}+}_i) \cdot U^{d/a}_{+}(i) + (x^{\mathrm{s}-}_i - \phi^{\mathrm{s}-}_i) \cdot U^{d/a}_{-}(i) + (x^{\bar{\mathrm{s}}}_i - \phi^{\bar{\mathrm{s}}}_i) \cdot U^{d/a}_{-}(i)]$

Now we will define the adversary observational expected utility of any signaling state $\hat{\omega} \in \Omega$. Let $U^{d/a}_{\sigma_j}(i, \eta^{\hat{\omega}}_i{=}1)$ be $U^{d/a}_{\sigma_j}(i)$ with $\eta^{\hat{\omega}}_i{=}1$ and $\eta^{\hat{\omega}'}_i = 0\ \forall \hat{\omega} \neq \hat{\omega}'$ and $j \in \{0, 1\}$.

4. $U^{d/a}_{\hat{\omega}}(\psi_i, \phi_i, x_i) = \Pr[\hat{\omega}|\mathrm{n}] \cdot [x^{\mathrm{n}-}_i \cdot U^{d/a}_{-}(i) + x^{\mathrm{n}+}_i \cdot U^{d/a}_{+}(i)] + \Pr[\hat{\omega}|\sigma_0] \cdot U^{d/a}_{\sigma_0}(i, \eta^{\hat{\omega}}_i{=}1) + \Pr[\hat{\omega}|\sigma_1] \cdot U^{d/a}_{\sigma_1}(i, \eta^{\hat{\omega}}_i{=}1)$ is the adversary observational expected utility. This is used in (A.5) and (A.10).

The set of constraints (A.2) - (A.4) enforce the randomized allocation is feasible, as in the version with only detection uncertainty. The set of constraints (A.5) - (A.7) ensure target t is the adversary's best response. b variables ensure adversary's utilities are nonnegative. The set of constraints (A.8)-(A.9) ensure the marginal probabilities of signaling (ψ and ϕ) are valid. Lastly, constraint (A.10) ensures η is a valid adversary behavior. In other words, if $\eta^{\hat{\omega}}_t$ is zero, then the adversary observational expected utility should be negative, otherwise the adversary utility should be positive.

A.3 Experimental Results

In Fig. 1.3e, we show the probability of a fake signal given that a warning signal is used. Equation A.11 describes this fully. This is then averaged over all of the targets, and finally averaged over 20 random graphs, as summarized in Equation A.12.

Plot	p-values
1.3a	$p \leq 3.457\text{e}{-16}$ for $N \geq 14$
1.3b	$p \leq 2.579\text{e}{-3}$ at $40 \leq N \leq 90$
1.3c	$p \leq 1.421\text{e}{-03}$ for $\gamma \geq 0.2$
1.3d	$p \leq 0.058$ for $\gamma \geq 0.4$
1.3e	$p = 2.167\text{e}{-22}$ when comparing $\gamma = 0$ and $\gamma = 0.9$
1.3f	1 vs. 2: $p \leq 1.371\text{e}{-04}$ for $\gamma \leq 0.7$ 2 vs. 3: $p \leq 2.852\text{e}{-02}$ for $\gamma \leq 0.7$ 1 vs. 3: $p \leq 1.984\text{e}{-05}$ for $\gamma \leq 0.8$
1.3g	$p \leq 6.661\text{e}{-02}$ at $\gamma = 0.3$ No difference at $\gamma = 0.5$ $p \leq 1.727\text{e}{-07}$ at $\gamma = 0.8$

Table A.1: p-values for results in Fig. 1.3 in the main paper.

$$P(fakesignal|\sigma_1)(i) = \frac{\gamma \cdot [(x_i^{s+} - \phi_i^{s+}) + (x_i^{s-} - \phi_i^{s-}) + (x_i^{\bar{s}} - \phi_i^{\bar{s}})]}{(1-\gamma) \cdot [(x_i^{s+} - \psi_i^{s+}) + (x_i^{s-} - \psi_i^{s-}) + (x_i^{\bar{s}} - \psi_i^{\bar{s}})] + \gamma \cdot [(x_i^{s+} - \phi_i^{s+}) + (x_i^{s-} - \phi_i^{s-}) + (x_i^{\bar{s}} - \phi_i^{\bar{s}})]} \tag{A.11}$$

$$\frac{1}{G}\frac{1}{N}\sum_{j=1}^{G}\sum_{i=1}^{N} P(\text{fake signal}|\sigma_1)(i) \tag{A.12}$$

The p-values for the experimental results are summarized in Table A.1. Fig. 1.3h does not have a p-value because it is based solely on the graph illustrated in Fig. 1.4, and the utilities described in Section A.4.

A.4 Conservation Drones

A.4.1 Utilities

The utilities used for the experiment in Section 1.7 are included in Table A.2. We construct this payoff matrix to reflect the fact that the reward and penalty of the adversaries are impacted by the following features: number of animals, distance to various park features such as boundary, rivers, and roads (some of the features used in[105] to predict poaching activity), and price.

To arrive at specific values, we first chose locations of interest near the park boundary, rivers, and roads in a region of the park known for the presence of animals. The park and specific coordinates are withheld to protect wildlife. We then measured the distances from the locations of interest to the closest rivers and roads, and the park boundary. The locations were ranked for each of these distances (e.g., node 6 is closest to the river and node 9 is farthest from the river, whereas node 9 is closest to a road and node 4 is farthest from a road). Next, a weighted average of these ranks was taken for each node to estimate the attractiveness of the node to animals (e.g., elephants), with weights of 0.8, 0.1, and 0.1 for distance to river, boundary, and road, respectively, according to the intuition that water matters most to animals. This was ranked from 1 to 10, with 10 being the best node for animals, and this served as a proxy for the number of animals at that node (e.g., 10 animals at node 6). To determine the relative poaching attractiveness for the adversary, the weighted average rank was calculated from the number of animals, and the river, boundary, and road distances, with weights of 0.7, 0.05, 0.15, and 0.1, respectively. This is based on the intuition that animals are the most important factor, but ease of reaching the location and getting away quickly is also a factor (e.g., nodes 6 and 7 are most attractive, while node 9 is least attractive).

We finally take elephants as an example animal, and use the price of ivory (approximately \$40,000[48]), the approximate monetary benefit of ecotourism (\$1.6 million[239]), and an elephant poaching fine (\$20,000[275]) to assign values to each of the 10 nodes. The defender payoffs are related to the eco-

Node	U_{du}	U_{dc}	U_{au}	U_{ac}
0	-3200	29	120	-20
1	-12800	55	320	-20
2	-8000	42	240	-20
3	-6400	38	160	-20
4	-4800	33	80	-20
5	-11200	51	280	-20
6	-16000	64	400	-20
7	-14400	59	400	-20
8	-9600	46	200	-20
9	-1600	24	40	-20

Table A.2: Utilities for Fig. 1.3h and Fig. 1.4 in the main paper.

tourism benefits of elephants – if the node is uncovered, it is related to the full amount, whereas if it is covered, it is related to an amount for one day. The adversary payoffs are related to the price of ivory, the attractiveness of a target, and the fines associated with a covered target. Given historical data, it may be possible to learn these values in the future from historical data[106], or possibly from park ranger knowledge[119].

A.4.2 Other False Negative Rates

We include results in Fig. 1.3h for a single $\gamma = 0.3$, but the relationship varies with γ as we have seen in the rest of Fig. 1.3. We include several other examples here for $\gamma = 0$, $\gamma = 0.1$, $\gamma = 0.5$, $\gamma = 0.7$, and $\gamma = 0.9$ in Figs. A.3a, A.3b, A.3c, A.3d, A.3e, respectively. At low values of γ, there is a small gap between ignoring detection uncertainty and GUARDSS, as expected, which indicates that it may be acceptable to ignore detection uncertainty at that point. However, as γ increases, the gap becomes wider, and ignoring detection uncertainty even becomes worse than a random allocation. The gap between no drones and GUARDSS also decreases as γ increases, meaning it becomes less beneficial to use drones under higher uncertainty, as seen in Fig. 1.3g.

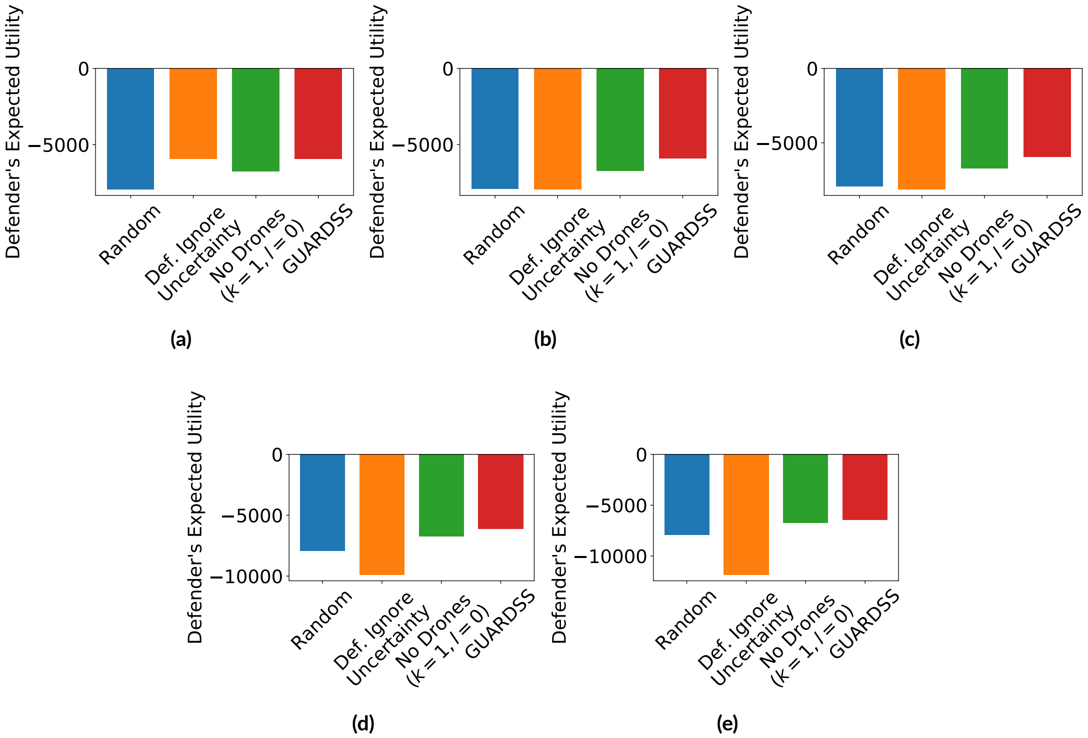


Figure A.3: Case study results for multiple values of γ: Fig. A.3a has $\gamma = 0$, Fig. A.3b has $\gamma = 0.1$, Fig.A.3c has $\gamma = 0.5$, Fig. A.3d has $\gamma = 0.7$, Fig. A.3e has $\gamma = 0.9$.

B

Appendix to Chapter 2

In this Appendix, we will provide additional details on the Serengeti dataset, choosing the deferral model, and further details regarding the human experiment and results.

B.1 Serengeti Dataset

B.1.1 Data Details

The Serengeti dataset is made up of animals, both adult and young, in the wild. Behaviors labeled in the dataset include standing, resting, moving, eating, and interacting. In addition, there may be some images with humans captured accidentally. We removed images that had been previously labeled to include humans, but there is a possibility that there are humans that were missed by previous labelers. For reference, the animal species captured include Grant's gazelles, reedbuck, dik dik, zebra, porcupine, Thomson's gazelles, spotted hyena, warthog, impala, elephant, giraffe, mongoose, buffalo, hartebeest, guinea fowl, wildebeest, leopard, ostrich, lion, kori bustard, other bird, bat eared fox, bushbuck, jackal, cheetah, eland, aardwolf, hippopotamus, striped hyena, aardvark, hare, baboon, vervet monkey, waterbuck, secretary bird, serval, topi, honey badger, rodents, wildcat, civet, genet, caracal, rhinoceros, reptiles, zorilla.

There are a variable number of human labels per image, as labels are collected until consensus is reached. All images have at least 5 labels, so we sample 5 randomly for each image. From the sampled data, we calculate the mean Cohen's kappa value to be 0.886, meaning very high agreement on a scale of -1 to 1. In fact, individual humans achieve 0.973 accuracy compared to consensus. While consensus is not guaranteed to be correct, Swanson et al.[282] show that on a gold standard dataset in which experts and many crowdsourced contributors label images, there was 96% agreement. We note that because some images were labeled in groupings of three due to the capture pattern and assigned the same labels for all, there may be some instances where one of the images is blank as the animal moves out of frame. From a random sample of 400 images, only 3 were mislabeled in the training set as containing animal species when there did not appear to be any present. These were when, out of the three pictures taken, an animal moved out of the camera's field of view in (usually) the last image.

B.1.2 Deferral Model Details

Our deferral model's objective function maximizes accuracy. Specifically, we define this as a weighted combination of sensitivity, the accuracy based on ground truth positive examples, and specificity, the accuracy based on ground truth negative examples. The weights to get the standard measure of accuracy are typically the number of positive and negative examples, respectively, but we allow them to be tuned to achieve different tradeoffs if desired. In our case, we choose a model based on weighting sensitivity and specificity of the composite model equally in the objective function, at 0.5 each. Fig. 3 in the main paper is generated by modifying the penalty of withholding between -0.5 and -0.1, inclusive, by -0.1.

We choose the point which achieves a deferral rate of 0.01. At a deferral rate of 0.01, if one SD card containing 5k images is processed at a time, about 50 images are deferred to a human. With about 20 SD cards per month, this leads to about 1000 images for human review, compared to 109k. At about 5-10 seconds for the difficult images, and 1 second for the easy images, this means that we ask for a maximum of about 3 hours of human time, compared to 302 hours.

Using this model, out of about 150k images in the Serengeti test set, a total of 1297 images are deferred, with 603 images containing animals and 694 empty images. We find some degree of complementarity, in that model accuracy on non-deferred images is 0.978, but 0.577 on deferred images. Furthermore, we mostly defer on empty images, as humans tend to get images containing animals incorrect (in the 15990 cases out of about 150k in which they are incorrect).

B.2 Human Experiment Details

We provide details of the ethical review in the ethics statement of the main paper. Further details are included below.

Comparison	Mean Values	Statistics (all significant)
DO vs. NM	61.9, 58.4	$t(197) = 3.8, p < 0.001$
DO vs. model alone (chance)	61.9, 50	$t(197) = 15.4, p < 0.001$
deferral status vs. no deferral status	60.4, 57.4	$F(1, 197) = 18.9, p < 0.001$
prediction vs. no prediction	57.8, 60.2	$F(1, 197) = 9.07, p = 0.003$
Conformity score > 0	0.08, 0	$t = 5.22, p < 0.0001$
Conformity score for low $>$ high confidence	0.116, 0.045	$t = 2.54, p = 0.014$
PO vs. other conditions (model incorrect)	41.9, 50.6	$t(197) = 23.9, p < 0.001$
Model correct vs incorrect (ims in NM and DO)	66.3, 50.5	$F(1, 39) = 2.19, p < 0.05$
Agreement where model correct vs. incorrect	69.6, 44.9	$t = 2.82, p = 0.007$

Table B.1: Summary of statistical results from human subject experiments.

B.2.1 Eligibility Criteria

We required participants to be consenting adults over 18 years old with intermediate-advanced English skills and good physical and mental health. Participants were requested from the UK and the US to increase the likelihood of intermediate-advanced English skills. Tasks were offered to workers via Prolific, an online automated system. Participants could abandon the task at any time. Domain experts were recruited by email.

B.2.2 Survey

We provide a link to one randomized version of the survey: https://bit.ly/SPM-Survey-AAAI2022.

B.2.3 Additional Results Details

Finally, we summarize our statistical findings from the main paper in Table B.1, and provide the full ANOVA in Table B.2.

	F Value	Num DF	Den DF	Pr > F
deferral status	18.9446	1.0000	197.0000	0.0000
prediction	9.0790	1.0000	197.0000	0.0029
model correct	859.1269	1.0000	197.0000	0.0000
deferral status:prediction	0.7850	1.0000	197.0000	0.3767
deferral status:model correct	2.3091	1.0000	197.0000	0.1302
prediction:model correct	12.5286	1.0000	197.0000	0.0005
deferral status:prediction:model correct	54.7870	1.0000	197.0000	0.0000

Table B.2: Full 2x2x2 ANOVA Results.

C

Appendix to Chapter 6

C.1 Dataset Additional Details

C.1.1 Noise and Occlusion Annotations

We handled noise and occlusion labels through a mixture of manually identifying these situations and automatically processing existing labels. We automatically considered labels to be occluded/occluding when the IoU is greater than 0.3. We also automatically considered frames to be noisy if there were a few missing labels in an object track, and interpolated missing labels. We use interpolation because, particularly in the case of ghosting or motion blur, the true bounding box is difficult to pinpoint due to noise. We provide examples of noise and occlusion annotations from this process in Fig. C.1 and Fig. C.2, respectively. We used the red labels in each case to represent normal animal labels, while the blue labels (in the middle frames) represent the animals with noise or occlusion. The separate distinction allows these cases to be used or discarded as needed depending on the task, whether object detection, tracking, etc.

C.1.2 Simulated Data

We added a lion to the simulation. We used 38 (311 K) for temperature in summer, 39 (310 K) for temperature in winter[295], and 0.98 for emissivity[199]. Object IDs and species labels for all objects of interest in the simulation were collected by using individual segmentation IDs corresponding to the actor name for each object. Videos were generated by following objects of interest with various offsets (sometimes within videos to break the smooth motion), camera angles, seasonality, and altitudes. Finally, if there was a small object along the border of an image, it was removed if less than 100 pixels in area. There is no noise in these synthetic data.

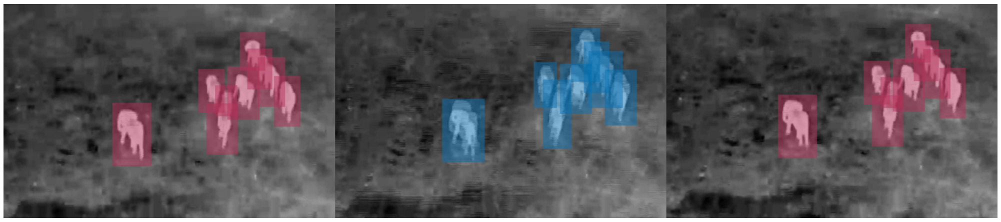


Figure C.1: Consecutive frames from a video in the dataset showing noise. Blue colored labels are noisy labels, while red are normal animal labels.

C.2 Additional Experiments

C.2.1 Detections

Table C.1 contains the results for two of the detection models on the proposed dataset, BIRDSAI, with ResNet[124] as the base model (instead of VGG16 results shown in Table 2 in the main paper), and the same experimental setup as described in Section 5.1 of the main paper. SSD and Faster-RCNN with ResNet perform better in some cases compared to VGG16, but overall, and especially for Faster-RCNN with weighted cross entropy, VGG16 outperforms ResNet.

Table C.2 (extension of Table 3 in the main paper, including the same Syn → Real row for easy reference) tabulates the performance baselines for detection in the unsupervised, semi-supervised and supervised domain adaptation setting. We still use the architecture from Domain Adaptive Faster-RCNN[64], but we include labeled real data at train time. The columns corresponding to FR-CE and FR-WCE are the standard Faster-RCNN trained over a training set that is a union of the synthetic and any available labeled real data (e.g., at 0% real data, it is only trained with synthetic data, while at 50% real data, all synthetic data is used plus half of the labeled real data). The columns for DA-FR-CE and DA-FR-WCE, on the other hand, indicate that in addition to the domain adaptive losses (image and instance level), the available labeled real data is also used to compute the label prediction loss included in the Domain Adaptive Faster-RCNN setting. We used three settings by

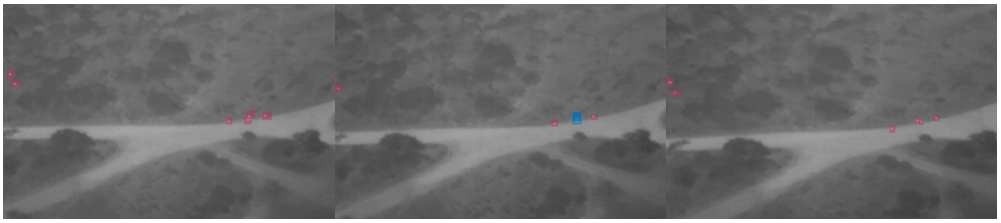


Figure C.2: Near-consecutive frames from a video in the dataset showing occlusion. Blue colored labels are occlusion labels, while red are normal animal labels.

Scale	FR-WCE (ResNet)	SSD (ResNet)
SA	0.202	0.137
MA	0.442	0.368
LA	0.884	0.886
Animals	**0.616**	**0.569**
SH	0.149	0.172
MH	0.193	0.214
LH	0.106	0.195
Humans	**0.142**	**0.196**
Overall	**0.403**	**0.390**

Table C.1: Detection performance baseline using the mAP metric for different scales ((S)mall, (M)edium, (L)arge) of objects ((A)nimals, (H)umans) in the dataset with ResNet as the base model.

Configuration	DA-FR-CE	DA-FR-WCE	FR-CE	FR-WCE
Syn → Real	0.443	0.459	0.309	0.313
Syn → Real (50% Sup. DA)	0.466	0.474	0.384	0.398
Syn → Real (100% Sup. DA)	0.522	0.518	0.448	0.472

Table C.2: Detection performance baselines using synthetic data. The mAP metric is reported.

using 0%, 50%, and 100% of the labeled real data to the training set of the synthetic data. All experiments were performed with VGG16 as the backbone network for 10 epochs with a batch size of 1, as well as an initial learning rate of 1e-4, a decay of 0.1 after a step of 4 epochs, and optimization with SGD. This table confirms that the synthetic data brings value despite the visible domain shift with respect to the real data. Unsupervised Domain Adaptation techniques help in improving performance, but using labeled real data improves the mAP results by over 10%. We expect BIRDSAI to be helpful in the development of more powerful unsupervised and semi-supervised domain adaptation techniques for object detection.

Species Recognition

We also annotated animal species in the real video frames where possible. The annotations were based on prior expert knowledge, as well as shape information. There were four different species

Species	Human	Elephant	Lion	Giraffe	Dog	Unknown
# bboxes	34001	83799	1244	12566	2709	21848
# frames	14959	13349	792	2242	2709	6804
mAP	0.068	0.305	0.004	0.142	0.002	0.237

Table C.3: Species label statistics and detection performance with Faster-RCNN on the real videos. The reported mAP values are computed over the test set.

apart from humans in the real dataset, and one label for *unknown*. We used these data to train Faster-RCNN (without weighting) with a total of six different classes. The annotation statistics and the test mAP are reported in Table C.3, with performance being loosely related to the number of examples and typical size of objects (e.g., there are many elephant examples, and these are typically large). There is room for improvement in all cases.

C.2.2 Tracking

Single Object Tracking

The comparison of SOT performance on *perfect subsequences* and *full sequences* (defined in Sec. 5.2 in the main paper) is included again in Table C.4 across different tracking algorithms - ADNet[*], ECO[†], Siamese RPN[‡] and MCFTS[§] - for ease. We show single object tracking (SOT) performance over the perfect subsequences and full sequences using the standard tracking metrics in Fig. C.3 and Fig. C.4, respectively.

We observe that Siamese RPN[175] performs very poorly on SOT in BIRDSAI. The Siamese RPN has been shown to work well in the visible spectrum and relies on visual one-shot detection in the current frame using an exemplar template. This approach seems to work poorly in the BIRDSAI dataset, likely given the limited textural details and poor resolution in the images due to the ther-

[*] https://github.com/hellbell/ADNet

[†] https://github.com/martin-danelljan/ECO

[‡] https://github.com/songdejia/Siamese-RPN-pytorch

[§] https://github.com/QiaoLiuHit/MCFTS

Method	Perfect Subsequences		Full Sequence	
	Precision	**AUC**	**Precision**	**AUC**
ECO	**0.8103**	**0.5430**	**0.4842**	**0.2972**
AD-Net	0.8029	0.5331	0.4545	0.2546
MCFTS	0.7194	0.4946	0.3401	0.1886
Siamese RPN	0.0073	0.0093	0.0041	0.0048

Table C.4: Single Object Tracking Evaluation. Precision is at 20 pixels. "Perfect subsequences" excludes noisy/occluded frames, while "Full sequence" includes them.

mal infrared sensing modality, and the sometimes large camera motion. ECO[79] also relies on some appearance-based cues and correlation filtering. However, it additionally learns a compact Gaussian Mixture Model (GMM)-based generative model of the target object and captures a diverse set of representations. Like Siamese RPN, MCFTS[181] also relies on deep convolutional networks, but it performs much better than the Siamese RPN in all cases. Because MCFTS uses convolutional features from a pre-trained network to form an ensemble of correlational trackers, we conjecture that the ensemble-based approach helps improve performance for weak trackers. AD-Net[339] is trained using a reinforcement learning-based approach where a convolutional neural network is trained as the policy function. The state is comprised of the cropped bounding box-based region of interest from the previous frame and a historical sequence of actions, where the actions capture the motion of the object's bounding box, e.g., left, right, far right, scale up/down, etc. The performance improvements of AD-Net possibly arise from the fact that it uses a history of actions, which captures the object motion from the last several frames.

The trackers that perform well on the *perfect subsequences* deteriorate when tested on *full sequences*. This performance drop is evident from the success and precision plots in Figs. C.3 and C.4. In most real-world scenarios, the sequences will be affected by noise, occlusions, the object leaving the frame and other such interruptions.

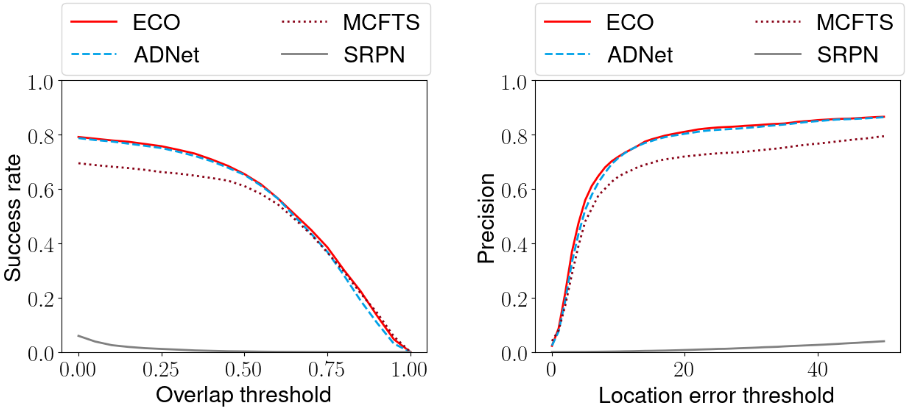


Figure C.3: Success and precision plots for the SOT with benchmark algorithms on *perfect subsequences*.

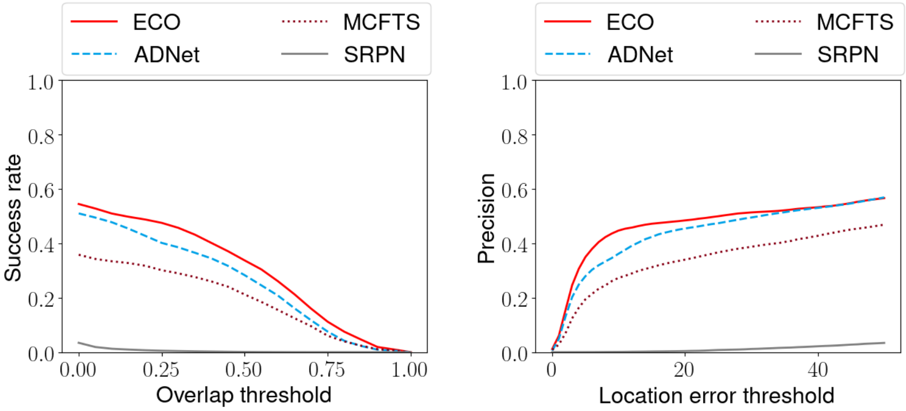


Figure C.4: Success and precision plots for the SOT with benchmark algorithms on the entire set of *full sequences*.

Method	**Object Size**	**MOTA**	**MOTP**
IoU Tracker (GT det.)	S	61.6	**100.0**
	M	**91.3**	98.9
	L	80.6	**100.0**
MDP Tracker (GT init.)	S	21.6	75.9
	M	54.6	84.1
	L	75.8	90.8

Table C.5: Multiple Object Tracking Evaluation. IoU tracker is given ground truth detections (GT det.), while an off-the-shelf MDP-based multi-object tracker is initialized using the ground truth detections (GT init.). S, M, L represents small, medium, and large objects, respectively.

Class	**FR-CE**	**FR-WCE**	**YOLOv2**	**SSD**	**DA-FRCE**	**DA-FRWCE**
Animals	0.188	0.204	0.074	0.058	0.112	0.117
Humans	0.177	0.186	0.032	0.092	0.107	0.142
Overall	0.181	**0.192**	0.044	0.089	0.110	0.129

Table C.6: Cross-Dataset Detection performance evaluation using the mAP metric.

Multi Object Tracking

Table C.5 tabulates the results obtained by trackers in the MOT setting, including results for IoU tracker provided in the main paper for easier reference. Off-the-shelf MDP[322] underperforms the IoU tracker, when the latter is provided with ground truth detections.

C.2.3 Cross-Dataset Evaluation

We also provide results trained using the LTIR dataset[27], as this was one of the most visually similar datasets to BIRDSAI. The results of cross-dataset detection on all of the baseline detectors as well as the domain adaptive detectors is shown in Table C.6. Based on these results, we conclude that the BIRDSAI dataset is substantially different than[27]. Moreover, based on the results in the previous sections, we can also conclude that it is sufficiently challenging by itself.

D

Appendix to Chapter 7

D.1 Additional Data Details

D.1.1 Satellite Data

We used the data shown in Table D.1. We acknowledge the use of data and/or imagery from NASA's Fire Information for Resource Management System (FIRMS) (`https://earthdata.nasa.gov/firms`), part of NASA's Earth Observing System Data and Information System (EOSDIS).

Feature	Collection Time	Google EE
Livestock Population Density[255]	2010	
Crop Cover[323]	2015	
Elevation[214]	2000	✓
Fire[107]	2016	✓
Fishing Hours[165]	2016	✓
Forest Cover[270]	2017	✓
Forest Change[121]	2017	✓
Landcover[53]	2017	✓
Nighttime Lights[89]	2017	✓
Population Density[68]	2015	✓
Presence of Water[235]	1984-2019	✓
Weather[200,213]	2017	✓
Crop Production[135]	2017	
Markets[110]	2017-2018	
Healthcare Sites[130]	2020	

Table D.1: Satellite data sources, collection time, and availability on Google Earth Engine.

D.1.2 MND Thresholds

In Table D.2, we include the thresholds used to define MND in this paper, though we are unable to provide raw data publicly.

MND	Biomarker	Values
Iron[a]	Ferritin	< 30 ng/mL
Vitamin A[b]	Retinol	< 0.20055 mg/L
Vitamin B12[c]	B12	< 300 pmol/L

Table D.2: Micronutrient deficiency thresholds.

[a] https://www.who.int/vmnis/indicators/serum_ferritin.pdf

[b] https://apps.who.int/iris/bitstream/handle/10665/44110/9789241598019_eng.pdf

[c] USDA

D.2 Additional Results

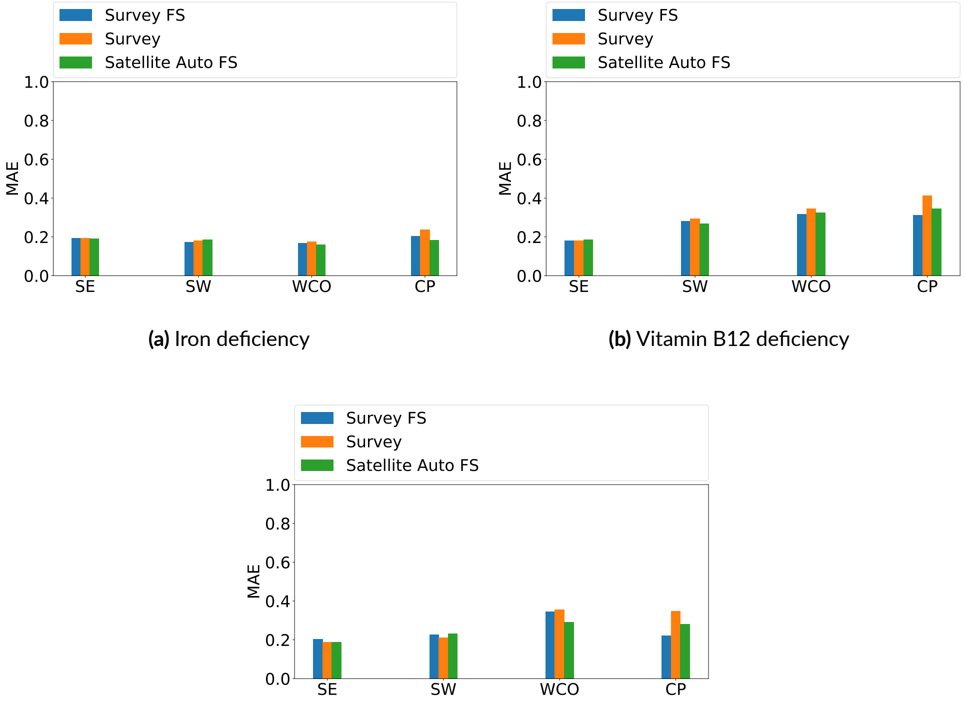
D.2.1 Socioeconomic Status

The causal mechanisms of MND are quite complex, but it is believed that there are multiple factors that influence MND. For example, we mention in "Possible Causes of MND" some environmental factors, especially forest presence, as well as socio-economic status. Epidemiological factors are also potential causes of MND, e.g., malaria.

In our analysis, we included multiple correlates to socio-economic status, such as nighttime lights, i.e., images of Earth at night, where it is expected that highly populous and resourced areas have more light. We found that the correlation coefficient between nighttime lights data and ground truth iron deficiency is 0.127, implying that alone, it may not be highly correlated. The individual feature with the greatest correlation is the sugarcane crop, with 0.147. If we predict solely with sugarcane, we achieve an AUC of 0.428, which is less than our findings of about 0.6 for iron deficiency. This implies that we need the other factors as well in order to predict MND.

D.2.2 Regression

We also report the regression results in Fig. D.1. We can see that the satellite imagery-based regression results are still comparable to the two versions of survey-based regression. In particular, MAE of our method ranges 0.16-0.19 in iron, 0.18-0.35 in Vitamin B12, and 0.19-0.29 in Vitamin A, which are reasonable considering the range of the regression task is $[0, 1]$ and the means are 0.21, 0.36, and 0.20, respectively. The AUC, F1-score, and MAE results all together demonstrate that our predictions are reasonably accurate.



(a) Iron deficiency

(b) Vitamin B12 deficiency

(c) Vitamin A deficiency

Figure D.1: Regression results comparison between satellite imagery based and survey based predictions. All elements are the same as Fig. 7.3 except that y-axis now means MAE of the regression task. Note that in this figure, *lower bars imply better results.*

D.2.3 Recall

Recall is important, as false negatives may lead to resources allocated away from people who truly have MND. Generally, recall is comparable to AUC for these data. However, it is higher in some cases. For example, for iron deficiency in region SE, recall is nearly 0.9.

References

[1] 932, T. F. A. (2013). Typical spectral response curves. http://flir.custhelp.com/app/answers/detail/a_id/932/~/typical-spectral-response-curves. Accessed: 2017-10-20.

[2] Abdur Rehman, N., Saif, U., & Chunara, R. (2019). Deep landscape features for improving vector-borne disease prediction. In *CVPR Workshops* (pp. 44–51).

[3] Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020). Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 252–260).

[4] Abebe, R. & Goldner, K. (2018). Mechanism design for social good. *AI Matters*, 4(3), 27–34.

[5] Air Shepherd (2019a). http://airshepherd.org.

[6] Air Shepherd (2019b). Air shepherd: The lindbergh foundation. http://airshepherd.org. Accessed: 2019-03-01.

[7] Alderson, D. L., Brown, G. G., Carlyle, W. M., & Wood, R. K. (2011). *Solving defender-attacker-defender models for infrastructure defense*. Technical report, Naval Postgraduate School.

[8] Alhussein, M. & Haider, S. I. (2016). Simulation and analysis of uncooled microbolometer for serial readout architecture. *Journal of Sensors*, 2016.

[9] Alpcan, T. & Basar, T. (2003). A game theoretic approach to decision and analysis in network intrusion detection. In *Proceedings of 42nd IEEE Conference on Decision and Control*, volume 3 (pp. 2595–2600).: IEEE.

[10] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–13).

[11] Anand, P., Hunter, G., Carter, I., Dowding, K., Guala, F., & Van Hees, M. (2009). The development of capability indicators. *Journal of Human Development and Capabilities*, 10(1), 125–152.

[12] Arp, G. & Phinney, D. (1980). The ecological variations in thermal infrared emissivity of vegetation. *Environmental and Experimental Botany*, 20(2), 135–148.

[13] Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of political economy*, 58(4), 328–346.

[14] Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, 58, 295–303.

[15] Ayush, K., Uzkent, B., Burke, M., Lobell, D., & Ermon, S. (2020). Generating interpretable poverty maps using object detection in satellite images. In *IJCAI* (pp. 4410–4416).

[16] Badrinarayanan, V., Galasso, F., & Cipolla, R. (2010). Label propagation in video sequences. In *CVPR* (pp. 3265–3272).: IEEE.

[17] Bae, S.-H. & Yoon, K.-J. (2014). Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *CVPR*.

[18] Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, 329(5995), 1081–1085.

[19] Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 (pp. 11405–11414).

[20] Basilico, N., De Nittis, G., & Gatti, N. (2015). A security game model for environment protection in the presence of an alarm system. In *GameSec*.

[21] Basilico, N., De Nittis, G., & Gatti, N. (2016a). A security game combining patrolling and alarm-triggered responses under spatial and detection uncertainties. In *AAAI*.

[22] Basilico, N., De Nittis, G., & Gatti, N. (2017). Adversarial patrolling with spatially uncertain alarm signals. *Artificial Intelligence*.

[23] Basilico, N., Nittis, G. D., & Gatti, N. (2016b). A security game combining patrolling and alarm–triggered responses under spatial and detection uncertainties. In *AAAI* (pp. 397–403).

[24] Bazilinskyy, P., Heisterkamp, N., Luik, P., Klevering, S., Haddou, A., Zult, M., Dialynas, G., Dodou, D., & de Winter, J. (2018). Eye movements while cycling in gta v. *Tools and Methods of Competitive Engineering*.

[25] Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in terra incognita. In *The European Conference on Computer Vision (ECCV)*.

[26] Beery, S., Wu, G., Rathod, V., Votel, R., & Huang, J. (2020). Context r-cnn: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13075–13085).

[27] Berg, A., Ahlberg, J., & Felsberg, M. (2015). A thermal object tracking benchmark. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–6).

[28] Berg, A., Ahlberg, J., & Felsberg, M. (2018). Generating visible spectrum images from thermal infrared. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[29] Berk, A., Bernstein, L. S., & Robertson, D. C. (1987). *MODTRAN: A moderate resolution model for LOWTRAN*. Technical report, SPECTRAL SCIENCES INC BURLINGTON MA.

[30] Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 401–413).

[31] Birhane, A. & Guest, O. (2020). Towards decolonising computational sciences. *Women, Gender and Research*.

[32] Black, E., Williams, J., Madaio, M. A., & Donti, P. L. (2020). A call for universities to develop requirements for community engagement in ai research. In *The Fair and Responsible AI Workshop at the 2020 CHI Conference on Human Factors in Computing Systems*.

[33] Blumenstock, J. (2018). Don't forget people in the use of big data for development. *Nature*, 561, 170–172.

[34] Bochinski, E., Senst, T., & Sikora, T. (2018). Extending iou based multi-object tracking by visual information. In *IEEE International Conference on Advanced Video and Signals-based Surveillance* (pp. 441–446). Auckland, New Zealand.

[35] Bogomolnaia, A., Moulin, H., & Stong, R. (2005). Collective choice under dichotomous preferences. *Journal of Economic Theory*, 122(2), 165–184.

[36] Bondi, E., Chen, H., Golden, C., Behari, N., & Tambe, M. (2022a). Micronutrient deficiency prediction via publicly available satellite data. In *IAAI*.

[37] Bondi, E., Dey, D., Kapoor, A., Piavis, J., Shah, S., Fang, F., Dilkina, B., Hannaford, R., Iyer, A., Joppa, L., & Tambe, M. (2018a). Airsim-w: A simulation environment for wildlife conservation with uavs. In *ACM COMPASS*.

[38] Bondi, E., Fang, F., Hamilton, M., Kar, D., Dmello, D., Choi, J., Hannaford, R., Iyer, A., Joppa, L., Tambe, M., & Nevatia, R. (2018b). Spot poachers in action: Augmenting conservation drones with automatic detection in near real time. In *IAAI*.

[39] Bondi, E., Fang, F., Kar, D., Noronha, V., Dmello, D., Tambe, M., Iyer, A., & Hannaford, R. (2017). Viola: Video labeling application for security domains. In *GameSec*.

[40] Bondi, E., Jain, R., Aggrawal, P., Anand, S., Hannaford, R., Kapoor, A., Piavis, J., Shah, S., Joppa, L., Dilkina, B., et al. (2020a). Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. In *WACV*.

[41] Bondi, E., Koster, R., Sheahan, H., Chadwick, M., Bachrach, Y., Cemgil, T., Paquet, U., & Dvijotham, K. (2022b). Role of human-ai interaction in selective prediction. In *AAAI*.

[42] Bondi, E., Oh, H., Xu, H., Fang, F., Dilkina, B., & Tambe, M. (2020b). To signal or not to signal: Exploiting uncertain real-time information in signaling games for security and sustainability. In *AAAI*.

[43] Bondi, E., Xu, L., Acosta-Navas, D., & Killian, J. A. (2021). Envisioning communities: A participatory approach towards ai for social good. In *AIES*.

[44] Boon, M., P. Drijfhout, A., & Tesfamichael, S. (2017). Comparison of a fixed-wing and multi-rotor uav for environmental mapping applications: A case study. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W6, 47–54.

[45] Boorman, E., O'Doherty, J., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, 80(6), 1558–1571.

[46] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[47] Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (2016). *Handbook of computational social choice*. Cambridge University Press.

[48] Brito, C. (2019). Ivory from hundreds of elephants found in $48 million seizure. *CBS News*.

[49] Brown, G., Carlyle, M., Salmerón, J., & Wood, K. (2006). Defending critical infrastructure. *Interfaces*.

[50] Brown, M. E., Grace, K., Shively, G., Johnson, K. B., & Carroll, M. (2014). Using satellite remote sensing and household survey data to assess human health and nutrition response to environmental change. *Population and environment*, 36(1), 48–72.

[51] Brown, N. & Sandholm, T. (2017). Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*.

[52] Bucarey, V., Casorrán, C., Figueroa, Ó., Rosas, K., Navarrete, H., & Ordóñez, F. (2017). Building real stackelberg security games for border patrols. In *GameSec*.

[53] Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., & Smets, B. (2020). Copernicus global land cover layers—collection 2. *Remote Sensing*, 12(6), 1044.

[54] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21.

[55] Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).: PMLR.

[56] CAASI (2021). Center for analytical approaches to social innovation.

[57] Catania, C. A., Bromberg, F., & Garino, C. G. (2012). An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection. *Expert Systems with Applications*, 39(2), 1822–1829.

[58] Cermak, J., Bosansky, B., Durkota, K., Lisy, V., & Kiekintveld, C. (2016). Using correlated strategies for computing stackelberg equilibria in extensive-form games. In *AAAI*.

[59] Černỳ, J., Boŷanskỳ, B., & Kiekintveld, C. (2018). Incremental Strategy Generation for Stackelberg Equilibria in Extensive-Form Games. In *EC*.

[60] Chajewska, U., Koller, D., & Parr, R. (2000). Making rational decisions using adaptive utility elicitation. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00)* (pp. 363–369).

[61] Chan, A. (2020). Approaching ethical impacts in ai research: The role of a graduate student. In *Resistance AI Workshop at NeurIPS*.

[62] Chavali, A. (2011). World health partners. *Hyderabad, India: ACCESS Health International Centre for Emerging Markets Solutions Indian School of Business*.

[63] Cheema, G. S. & Anand, S. (2017). Automatic Detection and Recognition of Individuals in Patterned Species. In *ECML PKDD*.

[64] Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster r-cnn for object detection in the wild. In *Computer Vision and Pattern Recognition (CVPR)*.

[65] Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1), 41–46.

[66] Christian Herrmann, Miriam Ruf, J. B. (2018). Cnn-based thermal infrared person detection by domain adaptation. In *Proc. SPIE 10643, Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, volume 10643.

[67] Christiansen, P., Steen, K. A., Jørgensen, R. N., & Karstoft, H. (2014). Automated detection and recognition of wildlife using thermal cameras. *Sensors*, 14(8), 13778–13793.

[68] CIESIN (2017). Gridded population of the world, version 4 (gpwv4): Population density, revision 11.

[69] Clark, D. A. (2005). Sen's capability approach and the many spaces of human well-being. *The Journal of Development Studies*, 41(8), 1339–1368.

[70] Coeckelbergh, M. (2010). Health care, capabilities, and ai assistive technologies. *Ethical theory and moral practice*, 13(2), 181–190.

[71] Cohen, J. E. (2012). *Configuring the networked self: Law, code, and the play of everyday practice*. Yale University Press.

[72] Conitzer, V. & Sandholm, T. (2006). Computing the Optimal Strategy to Commit to. In *EC*.

[73] Constanza-Chock, S. (2020). Introduction: #travelingwhiletrans, design justice, and escape from the matrix of domination. In *Design Justice*. The MIT Press, 1 edition. https://design-justice.pubpub.org/pub/ap8rgw5e.

[74] Cooney, S., Wang, K., Bondi, E., Nguyen, T. H., Vayanos, P., Winetrobe, H., Cranford, E. A., Gonzalez, C., Lebiere, C., & Tambe, M. (2019). Learning to signal in the goldilocks zone: Improving adversary compliance in security games. In *ECML/PKDD*.

[75] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016a). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[76] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016b). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[77] Crall, J., Stewart, C., Berger-Wolf, T., Rubenstein, D., & Sundaresan, S. (2013). Hotspotter – patterned species instance recognition. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on* (pp. 230–237).

[78] Critchlow, R., Plumptre, A. J., Driciru, M., Rwetsiba, A., Stokes, E. J., Tumwesigye, C., Wanyama, F., & Beale, C. (2015). Spatiotemporal trends of illegal activities from ranger-collected data in a ugandan national park. *Conservation Biology*, 29(5), 1458–1470.

[79] Danelljan, M., Bhat, G., Shahbaz Khan, F., & Felsberg, M. (2017). Eco: Efficient convolution operators for tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[80] De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12).

[81] De-Arteaga, M., Herlands, W., Neill, D. B., & Dubrawski, A. (2018). Machine learning for the developing world. *ACM Transactions on Management Information Systems (TMIS)*, 9(2), 1–14.

[82] de Cote, E., Stranders, R., Basilico, N., Gatti, N., & Jennings, N. (2013). Introducing alarms in adversarial patrolling games. In *AAMAS*.

[83] De Nittis, G. & Gatti, N. (2018). Facing Multiple Attacks in Adversarial Patrolling Games with Alarmed Targets. *arXiv preprint arXiv:1806.07111*.

[84] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR* (pp. 248–255).

[85] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114.

[86] D'Ignazio, C. & Klein, L. F. (2020). *Data feminism*. MIT Press.

[87] Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., & Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. In *The European Conference on Computer Vision (ECCV)*.

[88] Dulski, R., Madura, H., Piatkowski, T., & Sosnowski, T. (2007). Analysis of a thermal scene using computer simulations. *Infrared Physics & Technology*, 49(3), 257–260.

[89] Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C., & Ghosh, T. (2017). Viirs night-time lights. *International Journal of Remote Sensing*, 38(21), 5860–5879.

[90] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2), 303–338.

[91] Ewing-Nelson, C. (2021). All of the jobs lost in december were women's jobs. *National Women's Law Center Fact Sheet*.

[92] Fan, H. & Ling, H. (2017). Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. *arXiv preprint arXiv:1708.00153*.

[93] Fang, F., Nguyen, T. H., Pickles, R., Lam, W. Y., Clements, G. R., An, B., Singh, A., Tambe, M., & Lemieux, A. (2016). Deploying paws: Field optimization of the protection assistant for wildlife security. In *AAAI* (pp. 3966–3973).

[94] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.

[95] Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design ai for social good: Seven essential factors. *Science and Engineering Ethics*, 26(3), 1771–1796.

[96] Freytag, A., Rodner, E., Simon, M., Loos, A., Kühl, H. S., & Denzler, J. (2016). Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition* (pp. 51–63).: Springer.

[97] Gadiraju, K. K., Ramachandra, B., Chen, Z., & Vatsavai, R. R. (2020). Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery. In *KDD* (pp. 3234–3242).

[98] Gaidon, A., Wang, Q., Cabon, Y., & Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[99] Galesic, M. & Garcia-Retamero, R. (2010). Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples. *Archives of internal medicine*, 170(5), 462–468.

[100] Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1), 1–8.

[101] Gaver, B., Dunne, T., & Pacenti, E. (1999). Design: cultural probes. *interactions*, 6(1), 21–29.

[102] Geifman, Y. & El-Yaniv, R. (2017). Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*.

[103] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[104] Gholami, S., Ford, B., Fang, F., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., & Mabonga, J. (2017). Taking it for a test drive: A hybrid spatio-temporal model for wildlife poaching prediction evaluated through a controlled field test. In *ECML PKDD*.

[105] Gholami, S., Mc Carthy, S., Dilkina, B., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., Mabonga, J., et al. (2018). Adversary models account for imperfect crime data: Forecasting and planning against real-world poachers. In *AAMAS*.

[106] Gholami, S., Yadav, A., Tran-Thanh, L., Dilkina, B., & Tambe, M. (2019). Don't put all your strategies in one basket: Playing green security games with imperfect prior knowledge. In *AAMAS*.

[107] Giglio, L. & LANCE FIRMS (2016). Modis aqua & terra 1 km thermal anomalies and fire locations v006 nrt.

[108] Gillespie, D. T. (1996). The mathematics of brownian motion and johnson noise. *American Journal of Physics*, 64(3), 225–240.

[109] Girshick, R. (2015). Fast r-cnn. In *ICCV* (pp. 1440–1448).

[110] Golden, C. D., Rice, B. L., Randriamady, H. J., Vonona, A. M., Randrianasolo, J. F., Tafangy, A. N., Andrianantenaina, M. Y., Arisco, N. J., Emile, G. N., Lainandrasana, F., Mahonjolaza, R. F. F., Raelson, H. P., Rakotoarilalao, V. R., Rakotomalala, A. A. N. A., Rasamison, A. D., Mahery, R., Tantely, M. L., Girod, R., Annapragada, A., Wesolowski, A., Winter, A., Hartl, D. L., Hazen, J., & Metcalf, C. J. E. (2020). Study protocol: A cross-sectional examination of socio-demographic and ecological determinants of nutrition and disease across madagascar. *Frontiers in Public Health*, 8, 500.

[111] Gray, M. L. & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[112] Great Elephant Census (2016). The great elephant census, a paul g. allen project. Press Release.

[113] Green, B. (2018). Data science as political action: grounding data science in a politics of justice. *Available at SSRN 3658431*.

[114] Green, B. (2019). "Good" isn't good enough. In *AI for Social Good Workshop at NeurIPS*.

[115] Green, B. & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–24.

[116] Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*.

[117] Grote, T. (2021). Trustworthy medical ai systems need to know when they don't know. *Journal of Medical Ethics*, 47(5), 337–338.

[118] Guillory, D. (2020). Combating anti-Blackness in the AI community. *arXiv preprint arXiv:2006.16879*.

[119] Gurumurthy, S., Yu, L., Zhang, C., Jin, Y., Li, W., Zhang, X., & Fang, F. (2018). Exploiting data and human knowledge for predicting wildlife poaching. In *ACM COMPASS*.

[120] Hannaford, R. (2017). Eyespy. Private Communication.

[121] Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S., Goetz, S. J., Loveland, T. R., et al. (2013). High-resolution global maps of 21st-century forest cover change. *science*, 342(6160), 850–853.

[122] Haskell, W., Kar, D., Fang, F., Tambe, M., Cheung, S., & Denicola, E. (2014). Robust Protection of Fisheries with COmPASS. In *IAAI* (pp. 2978–2983).

[123] Hawken, P. (2017). *Drawdown: The most comprehensive plan ever proposed to reverse global warming*. Penguin.

[124] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

[125] He, Y., Ma, X., Luo, X., Li, J., Zhao, M., An, B., & Guan, X. (2017). Vehicle Traffic Driven Camera Placement for Better Metropolis Security Surveillance. *arXiv preprint arXiv:1705.08508*.

[126] Hill, M. (2003). Development as empowerment. *Feminist economics*, 9(2-3), 117–135.

[127] Hodgson, J. C., Baylis, S. M., Mott, R., Herrod, A., & Clarke, R. H. (2016). Precision wildlife monitoring using unmanned aerial vehicles. *Scientific reports*, 6.

[128] Hsiao, Y.-T., Lin, S.-Y., Tang, A., Narayanan, D., & Sarahe, C. (2018). vtaiwan: An empirical study of open consultation process in taiwan. *SocArXiv*.

[129] Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., & Smola, A. (2006). Correcting sample selection bias by unlabeled data. *NIPS*, 19, 601–608.

[130] Humanitarian Data Exchange (2020). Madagascar healthsites. https://data.humdata.org/dataset/madagascar-healthsites. Accessed: 2021-12-20.

[131] Hwang, S., Park, J., Kim, N., Choi, Y., & Kweon, I. S. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1037–1045).

[132] Ickowitz, A., Powell, B., Salim, M. A., & Sunderland, T. C. (2014). Dietary quality and tree cover in africa. *Global Environmental Change*, 24, 287–294.

[133] Ientilucci, E. J. & Brown, S. D. (2003). Advances in wide-area hyperspectral image simulation. In *Targets and Backgrounds IX: Characterization and Representation*, volume 5075 (pp. 110–122).: International Society for Optics and Photonics.

[134] Im, D. J., Kim, C. D., Jiang, H., & Memisevic, R. (2016). Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110*.

[135] International Food Policy Research Institute (2020). Spatially-Disaggregated Crop Production Statistics Data in Africa South of the Sahara for 2017.

[136] Ismail, A. & Kumar, N. (2021). Ai in global health: The view from the front lines. In *CHI Conference on Human Factors in Computing Systems*.

[137] Ivošević, B., Han, Y.-G., Cho, Y., & Kwon, O. (2015). The use of conservation drones in ecology and wildlife research. *Ecology and Environment*, 38(1), 113–188.

[138] Jain, M., Kardes, E., Kiekintveld, C., Ordónez, F., & Tambe, M. (2010). Security games with arbitrary schedules: A branch and price approach. In *AAAI*.

[139] Jeffries, B. (2017). CARE international annual report FY17. *CARE International*.

[140] Johnson, M. P., Fang, F., & Tambe, M. (2012). Patrol Strategies to Maximize Pristine Forest Area. In *AAAI*.

[141] Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S. N., Rosaen, K., & Vasudevan, R. (2017). Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *Robotics and Automation (ICRA), 2017 IEEE International Conference on* (pp. 746–753).: IEEE.

[142] Johnstone, J. (2007). Technology as empowerment: A capability approach to computer ethics. *Ethics and Information Technology*, 9(1), 73–87.

[143] Kahng, A., Lee, M. K., Noothigattu, R., Procaccia, A., & Psomas, C.-A. (2019). Statistical foundations of virtual democracy. In *International Conference on Machine Learning* (pp. 3173–3182).

[144] Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815), 169–169.

[145] Kamenica, E. & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*.

[146] Kamminga, J., Ayele, E., Meratnia, N., & Havinga, P. (2018). Poaching detection technologies—a survey. *Sensors*, 18(5).

[147] Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., & Wang, X. (2017). Object detection in videos with tubelet proposal networks. *arXiv preprint arXiv:1702.06355*.

[148] Kar, D., Fang, F., Fave, F. D., Sintov, N., & Tambe, M. (2015). "A Game of Thrones": When Human Behavior Models Compete in Repeated Stackelberg Security Games. In *AAMAS*.

[149] Kar, D., Ford, B., Gholami, S., Fang, F., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M., & Mabonga, J. (2017). Cloudy with a chance of poaching: Adversary behavior modeling and forecasting with real-world poaching data. In *AAMAS* (pp. 159–167).

[150] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

[151] Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., Bintz, C., Raz, D., & Krafft, P. (2020). Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 45–55).

[152] Keeley, B., Little, C., & Zuehlke, E. (2019). The state of the world's children 2019: Children, food and nutrition–growing well in a changing world. *UNICEF*.

[153] Kellenberger, B., Marcos, D., & Tuia, D. (2018). Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, 216, 139 – 153.

[154] Keswani, V., Lease, M., & Kenthapadi, K. (2021). Towards unbiased and accurate deferral to multiple experts. *arXiv preprint arXiv:2102.13004*.

[155] Kim, J.-H., Lim, S., Park, J., & Cho, H. (2019). Korean localization of visual question answering for blind people. In *AI for Social Good Workshop at NeurIPS*.

[156] Kleine, D. (2013). *Technologies of choice?: ICTs, development, and the capabilities approach*. MIT Press.

[157] Koenig, N. & Howard, A. (2004). Design and use paradigms for gazebo, an open-source multi-robot simulator. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2149–2154). Sendai, Japan.

[158] Koggel, C. (2003). Globalization and women's paid work: expanding freedom? *Feminist Economics*, 9(2-3), 163–184.

[159] Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1), 1–6.

[160] Kool, W. & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, 2(5995), 899–908.

[161] Koppmair, S., Kassie, M., & Qaim, M. (2017). Farm production, market access and dietary diversity in malawi. *Public health nutrition*, 20(2), 325–335.

[162] Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernández, G., Vojir, T., Hager, G., Nebehay, G., & Pflugfelder, R. (2015). The visual object tracking vot2015 challenge results. In *ICCV Workshops* (pp. 1–23).

[163] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

[164] Kroer, C., Waugh, K., Kilinc-Karzan, F., & Sandholm, T. (2017). Theoretical and practical advances on smoothing for extensive-form games. *EC*.

[165] Kroodsma, D. A., Mayorga, J., Hochberg, T., Miller, N. A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T. D., Block, B. A., et al. (2018). Tracking the global footprint of fisheries. *Science*, 359(6378), 904–908.

[166] Kuhlman, C., Jackson, L., & Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[167] Kumar, N., Karusala, N., Ismail, A., & Tuli, A. (2020). Taking the long, holistic, and intersectional view to women's wellbeing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(4), 1–32.

[168] Kumar, S. & Singh, S. K. (2017). Visual animal biometrics: survey. *IET Biometrics*, 6(3), 139–156.

[169] Kwok, R. (2019). Ai empowers conservation biology. *Nature*, 567(7746), 133–135.

[170] Latonero, M. (2019). Opinion: Ai for good is often bad. https://www.wired.com/story/opinion-ai-for-good-is-often-bad/.

[171] Lee, K.-H., Ros, G., Li, J., & Gaidon, A. (2019). SPIGAN: Privileged adversarial learning from simulation. In *International Conference on Learning Representations*.

[172] Lennon, J. J. (2000). Red-shifts and red herrings in geographical ecology. *Ecography*, 23(1), 101–113.

[173] Lentz, W. A. (1998). *Characterization of noise in uncooled IR bolometer arrays*. PhD thesis, Massachusetts Institute of Technology.

[174] Lewis, J. E., Abdilla, A., Arista, N., Baker, K., Benesiinaabandan, S., Brown, M., Cheung, M., Coleman, M., Cordes, A., Davison, J., et al. (2020). Indigenous protocol and artificial intelligence position paper. *Concordia Spectrum*.

[175] Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018a). High performance visual tracking with siamese region proposal network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[176] Li, C., Liang, X., Lu, Y., Zhao, N., & Tang, J. (2018b). RGB-T object tracking: Benchmark and baseline. *CoRR*, abs/1805.08982.

[177] Li, S. & Yeung, D.-Y. (2017). Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *AAAI*.

[178] Liao, Q. V. & Muller, M. (2019). Enabling value sensitive ai systems through participatory design fictions. *arXiv preprint arXiv:1912.07381*.

[179] Liew, A., Seong Cheng, W., Kumar, S., Neo, W., Tay, Y., Ng, K., & Lewis, J. (Accessed 2021). The pulse of the people. https://www.oppi.live/.

[180] Liu, Q. & He, Z. (2018). PTB-TIR: A thermal infrared pedestrian tracking benchmark. *CoRR*, abs/1801.05944.

[181] Liu, Q., Lu, X., He, Z., Zhang, C., & Chen, W.-S. (2017). Deep convolutional neural networks for thermal infrared object tracking. *Knowledge-Based Systems*, 134, 189 – 198.

[182] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., & Berg, A. C. (2015). SSD: single shot multibox detector. *CoRR*, abs/1512.02325.

[183] Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103.

[184] Lorde, A. (1984). The master's tools will never dismantle the master's house. In *Sister outsider: Essays and speeches*. Ten Speed Press.

[185] Luan, F., Paris, S., Shechtman, E., & Bala, K. (2017). Deep photo style transfer. *arXiv preprint arXiv:1703.07511*.

[186] Lucas, B. D., Kanade, T., et al. (1981). An iterative image registration technique with an application to stereo vision. *IJCAI*.

[187] Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS*, 30, 4765–4774.

[188] Lyu, S., Chang, M.-C., Du, D., Wen, L., Qi, H., Li, Y., Wei, Y., Ke, L., Hu, T., Del Coco, M., et al. (2017). Ua-detrac 2017: Report of avss2017 & iwt4s challenge on advanced traffic monitoring. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on* (pp. 1–7).: IEEE.

[189] Ma, C., Yang, X., Zhang, C., & Yang, M.-H. (2015). Long-term correlation tracking. In *CVPR* (pp. 5388–5396).

[190] Ma, Y., Wu, X., Yu, G., Xu, Y., & Wang, Y. (2016). Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery. *Sensors*, 16(4), 446.

[191] Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

[192] Madras, D., Pitassi, T., & Zemel, R. (2017). Predict responsibly: improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664*.

[193] Magassa, L., Young, M., & Friedman, B. (2017). Diverse voices: a method for underrepresented voices to strengthen tech policy documents. https://techpolicylab.uw.edu/wp-content/uploads/2017/10/TPL_Diverse_Voices_How-To_Guide_2017.pdf.

[194] MAHERY (2022). Mahery. https://www.mahery.org/. Accessed: 2022-05-11.

[195] Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. *Nature communications*, 9(1), 1–9.

[196] Malliaraki, E. (2019). What is this “ai for social good”? https://eirinimalliaraki.medium.com/what-is-this-ai-for-social-good-f37ad7ad7e91.

[197] Martin Jr., D., Prabhakaran, V., Kuhlberg, J., Smart, A., & Isaac, W. S. (2020). Participatory problem formulation for fairer machine learning through community based system dynamics. In *Eighth International Conference on Learning Representations (ICLR)*.

[198] Maskin, E. & Sen, A. (2014). *The Arrow impossibility theorem*. Columbia University Press.

[199] Mccafferty, D. J. (2007). The value of infrared thermography for research on mammals: previous applications and future directions. *Mammal Review*, 37(3), 207–223.

[200] McNally, A., Arsenault, K., Kumar, S., Shukla, S., Peterson, P., Wang, S., Funk, C., Peters-Lidard, C. D., & Verdin, J. P. (2017). A land data assimilation system for sub-saharan africa food and water security applications. *Scientific data*, 4(1), 1–19.

[201] Mhlambi, S. (2020). From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. *Carr Center Discussion Paper Series*.

[202] Micha, R., Mannar, V., Afshin, A., Allemandi, L., Baker, P., Battersby, J., Bhutta, Z., Chen, K., Corvalan, C., Di Cesare, M., et al. (2020). 2020 global nutrition report: action on equity to end malnutrition. *Development Initiatives*.

[203] Mikron Instrument Company, I. (n.d.). Table of emissivity of various surfaces. http://www.czlazio.com/tecnica/TabelladelleEmissivita.pdf. Accessed: 2022-04-23.

[204] Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

[205] Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, (pp. 1–26).

[206] Moore, J. (2019). Ai for not bad. *Frontiers in Big Data*, 2, 32.

[207] Moravčik, M., Schmid, M., Burch, N., Lisy, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., & Bowling, M. (2017). Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*.

[208] Mozannar, H. & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning* (pp. 7076–7087).: PMLR.

[209] Mueller, M., Smith, N., & Ghanem, B. (2016). A benchmark and simulator for uav tracking. In *Proc. of the European Conference on Computer Vision (ECCV)*.

[210] Müller, M., Casser, V., Lahoud, J., Smith, N., & Ghanem, B. (2018). Sim4cv: A photo-realistic simulator for computer vision applications. *International Journal of Computer Vision*, 126(9), 902–919.

[211] Nakalembe, C. (2020). Urgent and critical need for sub-saharan african countries to invest in earth observation-based agricultural early warning and monitoring systems. *Environmental Research Letters*, 15(12), 121002.

[212] Nardi, B., Tomlinson, B., Patterson, D. J., Chen, J., Pargman, D., Raghavan, B., & Penzenstadler, B. (2018). Computing within limits. *Communications of the ACM*, 61(10), 86–93.

[213] NASA GSFC HSL (2018). Fldas noah land surface model l4 global monthly 0.1 x 0.1 degree (merra-2 and chirps) v001.

[214] NASA JPL (2020). Nasadem merged dem global 1 arc second v001 [data set], nasa eosdis land processes daac. Accessed: 2020-12-30.

[215] News, B. I. (2021). New award supports social good. https://www.ischool.berkeley.edu/news/2021/new-award-supports-social-good.

[216] Nguyen, P., Kim, J., & Miller, R. C. (2013a). Generating annotations for how-to videos using crowdsourcing. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (pp. 835–840).

[217] Nguyen, T. H., Sinha, A., Gholami, S., Plumptre, A., Joppa, L., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Critchlow, R., et al. (2016). Capture: A new predictive anti-poaching tool for wildlife protection. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (pp. 767–775).: International Foundation for Autonomous Agents and Multiagent Systems.

[218] Nguyen, T. H., Yadav, A., An, B., Tambe, M., & Boutilier, C. (2014). Regret-Based Optimization and Preference Elicitation for Stackelberg Security Games with Uncertainty. In *AAAI*.

[219] Nguyen, T. H., Yang, R., Azaria, A., Kraus, S., & Tambe, M. (2013b). Analyzing the Effectiveness of Adversary Modeling in Security Games. In *AAAI* (pp. 718–724).

[220] Nguyen-Dinh, L.-V., Waldburger, C., Roggen, D., & Tröster, G. (2013). Tagging human activities in video by crowdsourcing. In *ICMR* (pp. 263–270).

[221] Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., & Clune, J. (2021). A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12(1), 150–161.

[222] Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25), E5716–E5725.

[223] Nussbaum, M. C. (2000). Introduction: Feminism and International Development. In *Women and human development: the capabilities approach*. Cambridge University Press.

[224] Nussbaum, M. C. (2011). *Creating capabilities*. Harvard University Press.

[225] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

[226] Olivares-Mendez, M. A., Bissyandé, T. F., Somasundar, K., Klein, J., Voos, H., & Le Traon, Y. (2014). The noah project: Giving a chance to threatened species in africa with uavs. In T. F. Bissyandé & G. van Stam (Eds.), *e-Infrastructure and e-Services for Developing Countries* (pp. 198–208).

[227] Olivares-Mendez, M. A., Fu, C., Ludivig, P., Bissyandé, T. F., Kannan, S., Zurad, M., Annaiyan, A., Voos, H., & Campoy, P. (2015). Towards an autonomous vision-based unmanned aerial system against wildlife poachers. *Sensors*, 15(12), 31362–31391.

[228] Oosterlaken, I. (2015). *Technology and human development*. Routledge.

[229] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1), 62–66.

[230] Pai, C.-H., Lin, Y.-P., Medioni, G. G., & Hamza, R. R. (2007). Moving object detection on a runway prior to landing using an onboard infrared camera. In *CVPR* (pp. 1–8).: IEEE.

[231] Pain, R. & Francis, P. (2003). Reflections on participatory research. *Area*, 35(1), 46–54.

[232] Park, H.-S. & Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2), 3336–3341.

[233] Park, S., Mohammadi, G., Artstein, R., & Morency, L.-P. (2012). Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface. In *CrowdMM* (pp. 29–34).

[234] Parvin, N. & Pollock, A. (2020). Unintended by design: On the political uses of "unintended consequences". *Engaging Science, Technology, and Society*.

[235] Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633), 418–422.

[236] Perez, C. C. (2019). *Invisible women: Exposing data bias in a world designed for men*. Random House.

[237] Perrault, A., Fang, F., Sinha, A., & Tambe, M. (2020). Ai for social impact: Learning and planning in the data-to-deployment pipeline. *AI Magazine*, 41(4).

[238] Pita, J., Jain, M., Western, C., Portway, C., Tambe, M., Ordonez, F., Kraus, S., & Paruchuri, P. (2008). Deployed ARMOR protection: The application of a game theroetic model for security at the Los Angeles International Airport. In *AAMAS*.

[239] Platt, J. R. (2014). Elephants are worth 76 times more alive than dead: Report. *Scientific American*.

[240] Poortinga, A., Nguyen, Q., Tenneson, K., Troy, A., Saah, D., Bhandari, B., Ellenburg, W. L., Aekakkararungroj, A., Ha, L., Pham, H., et al. (2019). Linking earth observations for assessing the food security situation in vietnam: a landscape approach. *Frontiers in Environmental Science*, 7, 186.

[241] Porikli, F., Bremond, F., Dockstader, S. L., Ferryman, J., Hoogs, A., Lovell, B. C., Pankanti, S., Rinner, B., Tu, P., & Venetianer, P. L. (2013). Video surveillance: past, present, and now the future [dsp forum]. *IEEE Signal Processing Magazine*, 30(3), 190–198.

[242] Portmann, J., Lynen, S., Chli, M., & Siegwart, R. (2014). People detection and tracking from aerial thermal views. In *2014 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1794–1800).

[243] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[244] Rasolofoson, R. A., Hanauer, M. M., Pappinen, A., Fisher, B., & Ricketts, T. H. (2018). Impacts of forests on children's diet in rural areas across 27 developing countries. *Science advances*, 4(8), eaat2853.

[245] Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. In *CSCW*.

[246] Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

[247] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *CVPR* (pp. 779–788).

[248] Redmon, J. & Farhadi, A. (2016). Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*.

[249] Ren, S., He, K., Girshick, R., & Sun, J. (2015a). Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

[250] Ren, S., He, K., Girshick, R., & Sun, J. (2015b). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*.

[251] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *KDD* (pp. 1135–1144).

[252] Richter, S. R., Vineet, V., Roth, S., & Koltun, V. (2016). Playing for data: Ground truth from computer games. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS* (pp. 102–118).: Springer International Publishing.

[253] Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*.

[254] Robichaud, C. (2005). *With great power comes great responsibility: On the moral duties of the super-powerful and super-heroic*. Open Court.

[255] Robinson, T. P., Wint, G. W., Conchedda, G., Van Boeckel, T. P., Ercoli, V., Palamara, E., Cinardi, G., D'Aietti, L., Hay, S. I., & Gilbert, M. (2014). Mapping the global distribution of livestock. *PloS one*, 9(5), e96084.

[256] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *MICCAI* (pp. 234–241).

[257] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016a). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3234–3243).

[258] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016b). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[259] Rosenfeld, A., Maksimov, O., & Kraus, S. (2018). Optimal cruiser-drone traffic enforcement under energy limitation. In *IJCAI*.

[260] Rowe, M. F., Bakken, G. S., Ratliff, J. J., & Langman, V. A. (2013). Heat storage in asian elephants during submaximal exercise: behavioral regulation of thermoregulatory constraints on activity in endothermic gigantotherms. *Journal of Experimental Biology*, 216(10), 1774–1785.

[261] Roy, J. (2019). Engineering by the numbers. In *Engineering by the numbers* (pp. 1–40).: American Society for Engineering Education.

[262] Saxena, R. S., Panwar, A., Semwal, S., Rana, P., Gupta, S., & Bhan, R. (2012). Pspice circuit simulation of microbolometer infrared detectors with noise sources. *Infrared Physics & Technology*, 55(6), 527–532.

[263] Schott, J. R. (2007). *Remote sensing: the image chain approach*. Oxford University Press on Demand.

[264] Sen, A. (1999a). *Development as Freedom*. Alfred A. Knopf.

[265] Sen, A. (1999b). The possibility of social choice. *American economic review*, 89(3), 349–378.

[266] Sen, A. (2017). *Collective choice and social welfare*. Harvard University Press.

[267] Shah, S., Dey, D., Lovett, C., & Kapoor, A. (2017). Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*.

[268] Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD* (pp. 614–622).

[269] Shi, Z. R., Wang, C., & Fang, F. (2020). Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*.

[270] Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., & Lucas, R. (2014). New global forest/non-forest maps from alos palsar data (2007–2010). *Remote Sensing of environment*, 155, 13–31.

[271] Shoham, Y. & Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

[272] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[273] Simonsen, J. & Robertson, T. (2012). *Routledge international handbook of participatory design*. Routledge.

[274] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

[275] Siyabona Africa (Pty) Ltd (2017). Stiffer penalties for poaching in zimbabwe. http://www.krugerpark.co.za/krugerpark-times-e-3-stiffer-penalties-for-poaching-in-zimbabwe-25062.html. Accessed: 2019-08-27.

[276] Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423*.

[277] Stein, G. J. & Roy, N. (2017). Genesis-rt: Generating synthetic images for training secondary real-world tasks. *arXiv preprint arXiv:1710.04280*.

[278] Steyn, N. P., Nel, J. H., Nantel, G., Kennedy, G., & Labadarios, D. (2006). Food variety and dietary diversity scores in children: are they good indicators of dietary adequacy? *Public health nutrition*, 9(5), 644–650.

[279] Sun, B., Feng, J., & Saenko, K. (2016). Return of frustratingly easy domain adaptation. In *AAAI*, volume 30.

[280] Sunderland, T., O'Connor, A., et al. (2020). Forests and food security: a review. *CAB Reviews*, 15(019), 1–10.

[281] Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. (2015a). Data from: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna.

[282] Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. (2015b). Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1), 1–14.

[283] Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, 56(5), 44–54.

[284] Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., et al. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4), 585–590.

[285] Tambe, M. (1995). Recursive agent and agent-group tracking in a real-time dynamic environment. In *ICMAS* (pp. 368–375).

[286] Tambe, M. (2011a). *Security and game theory: algorithms, deployed systems, lessons learned*. Cambridge University Press.

[287] Tambe, M. (2011b). *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press.

[288] Tambe, M. & Rosenbloom, P. S. (1995). Resc: An approach for real-time, dynamic agent tracking. In *IJCAI*, volume 95 (pp. 103–110).

[289] Tang, A. (2020). Inside taiwan's new digital democracy. *The Economist*.

[290] Thakoor, O., Tambe, M., Vayanos, P., Xu, H., Kiekintveld, C., & Fang, F. (2019). Cyber camouflage games for strategic deception. In *GameSec* (pp. 525–541).: Springer.

[291] Thinyane, H. & Bhat, K. S. (2019). Apprise: Supporting the critical-agency of victims of human trafficking in thailand. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

[292] Timmons, M. (2013). *Moral theory: An introduction*. Rowman & Littlefield Publishers.

[293] Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C., Ezer, D., van der Haert, F. C., Mugisha, F., et al. (2020). Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1), 1–6.

[294] Tran, T., Pham, T., Carneiro, G., Palmer, L., & Reid, I. (2017). A bayesian data augmentation approach for learning deep models. In *Advances in Neural Information Processing Systems* (pp. 2794–2803).

[295] Trethowan, P., Fuller, A., Haw, A., Hart, T., Markham, A., Loveridge, A., Hetem, R., du Preez, B., & Macdonald, D. W. (2017). Getting to the core: Internal body temperatures

help reveal the ecological function and thermal implications of the lions' mane. *Ecology and evolution*, 7(1), 253–262.

[296] Try AI, I. (2022). Try ai: Building connections and exploring ai in society. https://www.try-ai.org/. Accessed: 2022-05-11.

[297] Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., & Chandraker, M. (2018). Learning to adapt structured output space for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[298] Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458.

[299] Ul Haq, M. (2003). The birth of the human development index. *Readings in human development*, (pp. 127–137).

[300] UNODC (2016). World wildlife crime report: Trafficking in protected species. https://www.unodc.org/documents/data-and-analysis/wildlife/World_Wildlife_Crime_Report_2016_final.pdf.

[301] Van de Griend, A., Owe, M., Groen, M., & Stoll, M. (1991). Measurement and spatial variation of thermal infrared surface emissivity in a savanna environment. *Water resources research*, 27(3), 371–379.

[302] van Gemert, J. C., Verschoor, C. R., Mettes, P., Epema, K., Koh, L. P., Wich, S., et al. (2014). Nature conservation drones for automatic localization and counting of animals. In *ECCV Workshops (1)* (pp. 255–270).

[303] Venugopal, A., Bondi, E., Kamarthi, H., Dholakia, K., Ravindran, B., & Tambe, M. (2021). Reinforcement learning for unified allocation and patrolling in signaling games with uncertainty. In *AAMAS*.

[304] von Grebmer, K., Saltzman, A., Birol, E., Wiesman, D., Prasai, N., Yin, S., Yohannes, Y., Menon, P., Thompson, J., Sonntag, A., et al. (2014). 2014 global hunger index: The challenge of hidden hunger. *IFPRI books*.

[305] Vondrick, C., Patterson, D., & Ramanan, D. (2011). Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, (pp. 1–21).

[306] Voss, R. F. (1978). Linearity of 1 f noise mechanisms. *Physical Review Letters*, 40(14), 913.

[307] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset*. Technical Report CNS-TR-2011-001, California Institute of Technology.

[308] Wang, B., Zhang, Y., & Zhong, S. (2017). On repeated stackelberg security game with the cooperative human behavior model for wildlife protection. In *AAMAS*.

[309] Wang, F. & Preininger, A. (2019). Ai in health: state of the art, challenges, and future directions. *Yearbook of medical informatics*, 28(1), 16.

[310] Wang, Y., Shi, Z. R., Yu, L., Wu, Y., Singh, R., Joppa, L., & Fang, F. (2019). Deep reinforcement learning for green security games with real-time information. In *AAAI*.

[311] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., & Perona, P. (2010). *Caltech-UCSD Birds 200*. Technical Report CNS-TR-2010-001, California Institute of Technology.

[312] Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M., et al. (2019). Disability, bias, and ai. *AI Now Institute*.

[313] Wiener, Y. & El-Yaniv, R. (2013). *Theoretical foundations of selective prediction*. PhD thesis, Computer Science Department, Technion.

[314] Wilder, B., Horvitz, E., & Kamar, E. (2020). Learning to complement humans. *arXiv preprint arXiv:2005.00582*.

[315] Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1), 80–91.

[316] Winner, L. (1980). Do artifacts have politics? *Daedalus*, (pp. 121–136).

[317] Witham, C. L. (2017). Automated face recognition of rhesus macaques. *Journal of neuroscience methods*.

[318] Wolff, J. & De-Shalit, A. (2007). *Disadvantage*. Oxford university press on demand.

[319] Wu, Y., Lim, J., & Yang, M.-H. (2015). Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9), 1834–1848.

[320] Wu, Z., Fuller, N., Theriault, D., & Betke, M. (2014). A thermal infrared video benchmark for visual analysis. In *2014 IEEE CVPR Workshop on Perception Beyond the Visible Spectrum* (pp. 201–208).

[321] Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2018). Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[322] Xiang, Y., Alahi, A., & Savarese, S. (2015). Learning to track: Online multi-object tracking by decision making. In *International Conference on Computer Vision (ICCV)*.

[323] Xiong, J., Thenkabail, P., Tilton, J., Gumma, M., Teluguntla, P., Congalton, R., Yadav, K., Dungan, J., Oliphant, A., Poehnelt, J., Smith, C., & Massey, R. (2017). Nasa making earth system data records for use in research environments (measures) global food security-support analysis data (gfsad) cropland extent 2015 africa 30 m v001.

[324] Xu, H., Ford, B., Fang, F., Dilkina, B., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., & Nsubaga, M. (2017). Optimal patrol planning for green security games with black-box attackers. In *GameSec*.

[325] Xu, H., Rabinovich, Z., Dughmi, S., & Tambe, M. (2015). Exploring Information Asymmetry in Two-Stage Security Games. In *AAAI* (pp. 1057–1063).

[326] Xu, H., Wang, K., Vayanos, P., & Tambe, M. (2018a). Strategic coordination of human patrollers and mobile sensors with signaling for security games. In *AAAI*.

[327] Xu, H., Wang, K., Vayanos, P., & Tambe, M. (2018b). Strategic coordination of human patrollers and mobile sensors with signaling for security games. In *AAAI*.

[328] Xu, L., Bondi, E., Fang, F., Perrault, A., Wang, K., & Tambe, M. (2021). Dual-mandate patrols: Multi-armed bandits for green security. In *AAAI*.

[329] Yahyanejad, S. & Rinner, B. (2015). A fast and mobile system for registration of low-altitude visual and thermal aerial images using multiple small-scale uavs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 104, 189–202.

[330] Yan, R., Yang, J., & Hauptmann, A. (2003). Automatically labeling video data using multi-class active learning. In *ICCV*.

[331] Yang, J., Lu, J., Batra, D., & Parikh, D. (2017). A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*.

[332] Yang, Q., Steinfeld, A., & Zimmerman, J. (2019). Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11).

[333] Yang, R., Kiekintveld, C., Ordonez, F., Tambe, M., & John, R. (2011). Improving resource allocation strategy against human adversaries in security games. In *IJCAI*.

[334] Yearsley, J. (2016). Generate spatial data. Accessed: 2018-03-01.

[335] Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4), 13.

[336] Yin, Y., An, B., & Jain, M. (2014). Game-theoretic resource allocation for protecting large public events. In *AAAI*.

[337] Yin, Z., Jain, M., Tambe, M., & Ordonez, F. (2011). Risk-averse strategies for security games with execution and observational uncertainty. In *AAAI*.

[338] Young, M., Magassa, L., & Friedman, B. (2019). Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, 21(2), 89–103.

[339] Yun, S., Choi, J., Yoo, Y., Yun, K., & Young Choi, J. (2017). Action-decision networks for visual tracking with deep reinforcement learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[340] Zhang, H. & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience*, 6, 1.

[341] Zhang, L., Li, Y., & Nevatia, R. (2008). Global data association for multi-object tracking using network flows. In *CVPR* (pp. 1–8).: IEEE.

[342] Zhang, Y., David, P., & Gong, B. (2017). Curriculum domain adaptation for semantic segmentation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2 (pp.6).

[343] Zhang, Y., Guo, Q., An, B., Tran-Thanh, L., & Jennings, N. R. (2019). Optimal interdiction of urban criminals with the aid of real-time information. In *AAAI*.

[344] Zhao, B., Feng, J., Wu, X., & Yan, S. (2017). A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2), 119–135.

[345] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*.

[346] Zhu, Y., Lan, Z., Newsam, S., & Hauptmann, A. G. (2017b). Guided optical flow learning. *arXiv preprint arXiv:1702.02295*.

[347] Zyskowski, K., Morris, M. R., Bigham, J. P., Gray, M. L., & Kane, S. K. (2015). Accessible crowdwork? understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1682–1693).