Student's Declaration

I hereby declare that the work presented in the report entitled **"Computer Vision Applications in Wildlife Conservation"** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Engineering* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Saket Anand**. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

..... **Place & Date: .....**
Palash Aggrawal

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

..... **Place & Date: .....**
Dr. Saket Anand

Abstract

Monitoring of protected areas to curb illegal activities like poaching and animal trafficking is a monumental task. To augment existing manual patrolling efforts, unmanned aerial surveillance using visible and thermal infrared (TIR) cameras is increasingly being adopted. Automated data acquisition has become easier with advances in unmanned aerial vehicles (UAVs) and sensors like TIR cameras, which allow surveillance at night when poaching typically occurs. However, it is still a challenge to accurately and quickly process large amounts of the resulting TIR data. In this paper, we present the first large dataset collected using a TIR camera mounted on a fixed-wing UAV in multiple African protected areas. This dataset includes TIR videos of humans and animals with several challenging scenarios like scale variations, background clutter due to thermal reflections, large camera rotations, and motion blur. We also evaluate various recent approaches for single and multi-object tracking. With the increasing popularity of aerial imagery for monitoring and surveillance purposes, we anticipate this unique dataset to be used to develop and evaluate techniques for object detection, tracking, and domain adaptation for aerial, TIR videos. To this end, we explore the use of Person Re-Identification(ReID) techniques for applicability in multi target multi camera tracking. We find that ReId is best suited for multi camera target reidentification compared to single camera tracking

Keywords: Computer Vision, Wildlife Conservation, Infrared tracking, UAV Tracking, Mutli Camera Tracking, MTMC, Target Reidentification, Tracking Dataset

Acknowledgments

I would like to express my gratitude to my advisor: Dr. Saket Anand, for giving me this opportunity, guiding me throughout, entrusting my capabilities and providing the much-needed insights and ideas. I would also like to thank Raghav Jain for working along with me through 2 semesters on the project.
I would like to express my special gratitude to Elizabeth Bondi (Harvard) and Anil Sharma (IIITD), along with all the co-authors of works I got a chance to collaborate in while working on this project. This research resulted in the following works of which I was a co-author:

- BIRDSAI: A Dataset for Detection and Tracking in Aerial Thermal Infrared Videos [8]
- Scalable Camera Selection Decisions in a Multi-Camera Network (submitted to ACM Multimedia)

This report highlights my contributions in them and, therefore, has material from both these works.

The BIRDSAI dataset was supported by Microsoft AI for Earth, NSF CCF-1522054 and IIS-1850477, MURI W911NF-17-1-0370, and the Infosys Center for Artificial Intelligence, IIIT-Delhi. I am very thankful to my institute, Indraprastha Institute of Information Technology, for supporting this research.

Contents

Chapter 1

Introduction

1.1 Surveillance for Wildlife Conservation

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 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 [20, 26, 28, 29, 33, 39, 41, 51, 77], there has been an increasing interest in applications from the near-infrared (NIR) and thermal infrared (TIR) spectral ranges [3, 35, 46, 49, 76], 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 [28, 47, 77]. In [77], 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 UAVs [28,47] are often motivated by applications like surveillance and 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.

Over the last few years, various technological interventions have surfaced, and have led to the

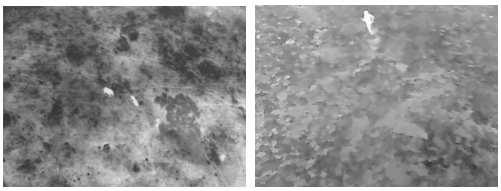


Figure 1.1: Example images from BIRDSAI: elephants and a human, respectively, from an aerial perspective.

Dataset (Year)	Platform (A/G)	#Frames	Tasks	Spectrum	(R)eal/(S)ynthetic
UTB [47] (2017)	A	15K	S	V	R
UAV123 [54] (2016)	A	113K	S	V	R,S
UAVDT [28] (2018)	A	80K	D,S,M	V	R
TIV [76] (2014)	G,A[a]	64K	D,S,M	T	R
LTIR [3] (2015)	G,A	12K	D,S,M	T	R
PTB-TIR [49] (2018)	G,A	30K	D,S,M	T	R
ASL-TID [56] (2014)	A[a]	5K	D,S,M	T	R
[79] (2015)	G,A[b]	84[c]	RE	T,V	R
[52][d] (2016)	A	9K	D,S,M	T	R
BIRDSAI	A	62K + 100K	D,S,M	T	R,S

Table 1.1: 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. 2.1. [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.

release of public datasets advancing wildlife monitoring applications like species categorization [2,66,73] and individual identification of chimpanzees [32] and rhesus macaques [74]. In addition to camera trapping and similar non-invasive terrestrial imaging techniques, aerial imagery has long been used for wildlife monitoring [34]. These publicly released datasets have had tremendous impact

Monitoring of animals addresses only one aspect of conservation planning and management. When trying to mitigate illegal activities like poaching, hunting and logging, surveillance efforts in protected areas at night-time is very challenging and puts forest rangers at risk due to poor visibility, difficult terrain and increased predator activity [55]. To curb poaching activities, UAV surveillance is increasingly becoming popular [1, 38]. TIR cameras serve as an effective sensing modality for night-time aerial surveillance over natural landscapes, where the ambient light is minimal and the UAV's altitude, payload capacity, and need for stealth restrict the use of active light sensors. As a result, manual monitoring of surveillance videos has been necessary, which is an extremely tedious task, especially with when the goal is near real-time response in order to interdict an illegal activity. The video monitoring problem is exacerbated by several challenges in using TIR cameras: low-contrast and noisy images, background clutter due to spurious thermal reflections from the ground and other irrelevant objects, and the camera motion due to the UAV's flight path. Automatic surveillance could directly reduce poaching incidents by detecting and tracking humans in restricted areas.

1.2 Multi Target Multi Camera Tracking using Re-identification

Camera networks are pervasive and frequently used for various visual analytics and multimedia applications in robot perception, video surveillance, crowd behavior analysis, etc. For example,

tracking vehicles across multiple cameras deployed on road intersections to estimate the drive time of the different road segments. The tracking task aims to determine the position of a target (person, vehicle, animal, etc.) at all times across the different video streams of the camera network. The number of cameras at an airport, train station, malls, etc. has rapidly increased, which makes automated tracking an essential task for visual analytics.

Usually, a multi-camera tracking algorithm makes a Re-ID query to identify the target's location in the camera network, which when made at all times in all cameras is likely to generate false alarms leading to loss of the target being tracked [60, 63]. Secondly, as most recent Re-ID techniques are deep learning based, a large number of queries also increases the computational cost manifold. This redundant querying has been addressed by modeling the transition time of the target using a static distribution like a Gaussian [44] or Parzen window based [36]. However, this transition time distribution need not be static, and may depend on characteristics of the target as well as the environment, e.g., target speed, congestion due to other objects present, slippery floor, etc. This observation motivates our strategy to model this transition time in a time-dependent manner by defining a state vector that encodes the target's location in the camera network. Later, we highlight that this state representation is crucial from a scalability perspective.

Many approaches for multi-camera target tracking employ a two-step framework [44, 60, 70, 78]. First, SCT (Single-Camera Tracking) to find the target's trajectory within each camera. Second, ICT (Inter-Camera Tracking), to associate the SCT trajectories corresponding to the same identity. ICT is used when the target is transitioning between different cameras or when it is occluded in a single camera. SCT is a relatively easier problem where the spatial correlation of the target's location between consecutive frames can be effectively leveraged. On the other hand, the ICT problem is harder due to the indeterminate transition time, and the existing solutions typically focus on Re-ID and data (or target trajectory) association. Instead of improving the quality of Re-ID or data association approaches, in this paper we focus on answering *when* should a Re-ID query be made, and to *which* camera frame. We formulate the camera selection decisions problem as a Markov Decision Process (MDP) and learn a policy using reinforcement learning (RL). The learning is directly from the visual camera feed and does not require the knowledge of the camera network topology.

1.3 Contributions

Motivated by the problems pointed out in 1.1, 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

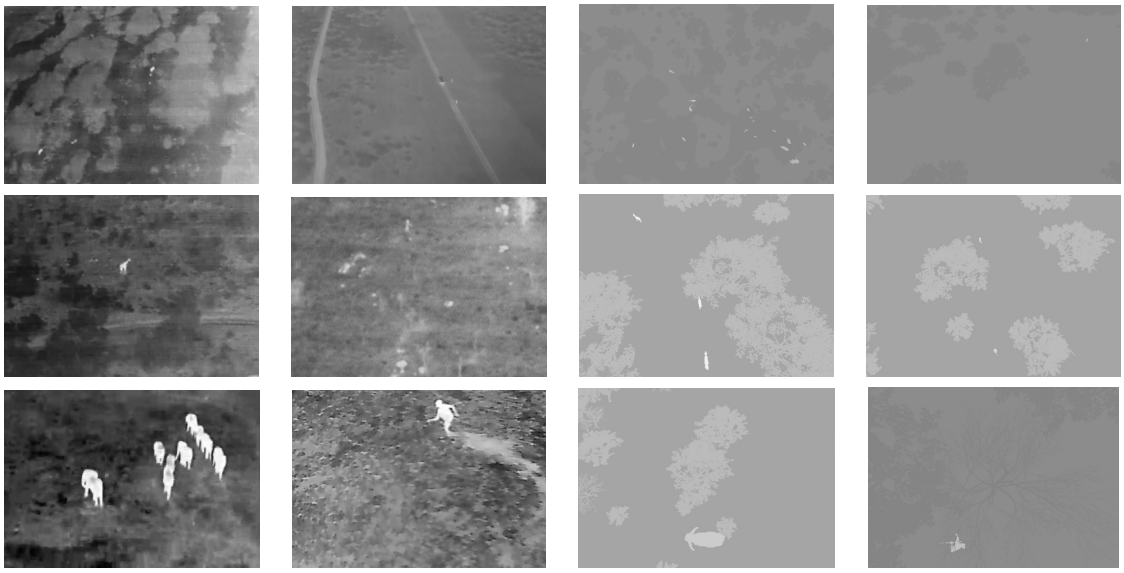
fixed-wing UAV over national parks in Southern Africa. Two example images from real videos are shown in Fig. 1.1 depicting a herd of elephants and a human. Realistic and challenging benchmarking datasets have had tremendous impact on the progress of a research area. The Caltech-UCSD Bird (CUB-200) dataset [71, 73] has helped advance an important area of fine-grained visual recognition [83]. With more wildlife monitoring datasets [2, 32, 66, 74] becoming publicly available, we may expect rapid progress in areas like species detection, counting, and visual animal biometrics [11, 21, 32, 42]. 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.

After creating the dataset and recognising the challenging environment it poses, we work towards using Re-Identification approach for identification of targets and single object as well as multi object tracking. We study Re-ID in a multi target multi camera (MTMC) tracking setting to understand its dymanics and performance. We start from pretrained state-of-the-art Re-ID network and also additionally by training it on challenging Re-ID datasets. We do extensive experimentation to find the optimal performance point for the network, thereby discovering the important hyperparameters to use for best results. We perform Inter Camera Tracking (ICT) and Single Camera Tracking (SCT) using Re-ID to study its performance and applicability. The Re-ID network is integrated in a larger Camera Selection Model to aid in its ICT and SCT tasks.

Chapter 2

Related Work and Motivation

2.1 A large UAV-TIR forest dataset



(a) (b) (c) (d)

Figure 2.1: 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).

With poaching becoming widespread around the world [69], aerial surveillance with UAVs is becoming a mainstream application [38, 40, 55]. In order to apply deep learning-based detection 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 UAVs. 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 1.1.

Existing UAV Datasets: The recently introduced UAVDT [28] 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 [47] was introduced for benchmarking UAV-based single object tracking with the goal of jointly evaluating the motion model and tracking performance. Mueller et al. introduced the UAV123 dataset [54], which contains 123 HD video sequences with about 113,000 annotated frames captured by a low-altitude UAV. Eight of these videos were rendered using an Unreal Engine environment. All of these datasets use visible spectrum cameras mounted on multirotor UAVs, which typically have lower speeds and better image stabilization as compared to fixed-wing UAVs [9]. In a poaching prevention application, deploying a multirotor UAV for surveillance is more difficult due to stealth and coverage requirements.

Existing TIR Datasets: The BU-TIV dataset [76] is part of the OTCBVS dataset collection[1] and contains 16 video sequences with over 60,000 annotated frames for tasks like detection, counting and tracking. The LTIR [3] dataset was used for the VOT-TIR 2016 challenge and contains 20 video sequences of length 563 frames on average. The PDT-ATV dataset [56] 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 UAV, but because it is handheld, it is a fixed aerial perspective. Recently, the PTB-TIR dataset [49] 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 UAV* has not been addressed yet.

BIRDSAI: The 48 real TIR video sequences included in BIRDSAI were randomly selected from a database of UAV 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.

2.2 Re-identification in a Mutli Camera Multi Target Tracking Setting

To understand and study the usability of Re-ID for the challenging BIRDSAI dataset, we start by studying it in the context of Multi Camera Multi Target (MTMC) tracking.

In a multi camera network, to resolve camera handovers in non-overlapping Field Of View (FOVs), a few initial works have created a social group model [82] to associate target tracklets, affinity model [43] of target's appearance for inter-camera association. Other works formulate various data association methods [16, 22, 53] to resolve camera handovers and use graph [12, 30,

[1] `http://vcipl-okstate.org/pbvs/bench/`

31, 48, 70, 72, 82] based approaches for inter-camera tracking. Spatio-temporal contextual information [78], clique based methods [57, 59], part based model [4, 65] are also a few other common approaches. Many work perform pairwise matching [10, 19, 24, 25, 62, 67] of the templates to form trajectories. Template re-identification [67, 80] approaches are leading for matching target's template with other candidate templates. To resolve the all pair matching issues, multiple method [17, 60] associate only time consecutive templates to reduce computational complexity [60]. In this regard, works [36, 53] use the travel time of the target to estimate the transition time of the camera handover. Works [44] have estimated a transition time distribution using a Gaussian distribution.

Association based works perform multi-camera target tracking in a unified way. Recent works for multi-camera target tracking perform tracking task in a two step framework. First, they perform single camera tracking (SCT) and then inter-camera tracking (ICT) to resolve the camera handover separately. Works such as [17, 44, 60, 68, 70, 78] use such a two step framework for tracking in multiple cameras. Current state-of-the-art in appearance features is in deep learning based methods to re-identify a target in different cameras through various viewpoints [37, 60, 80] including deep feature representation learning, deep metric learning and ranking optimization. [60] learn the correlation features using combinatorial optimization. They have proposed a weighted triplet loss to learn better features of target's appearance. However, their approach tracks a target in an offline fashion and makes a very large number of reid queries to the camera network. We use [13] in our work to re-identify a target in any camera.

Chapter 3

Methodology

3.1 The BIRDSAI Dataset

The BIRDSAI dataset is created from both Real and Synthetic domains, this report focuses on the Real data.

3.1.1 Dataset Description

Data Acquisition

Data were collected throughout protected areas in the countries of South Africa, Malawi, and Zimbabwe using a battery-powered fixed-wing UAV. 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.

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 [18]. 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

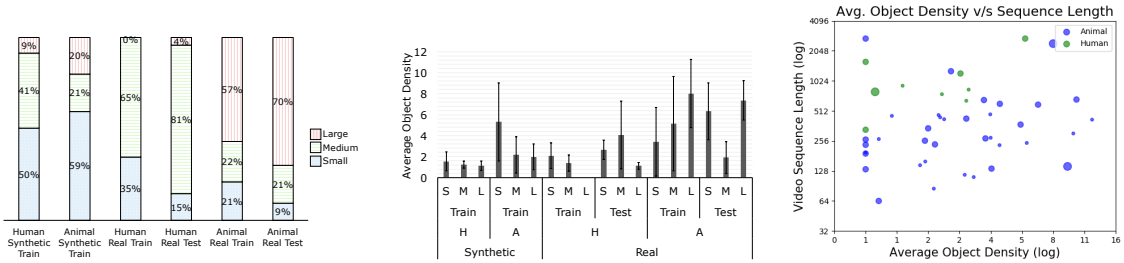
added robustness is crucial for building practical vision systems that are less sensitive to specific UAV or camera settings.

Annotation

We used VIOLA [7] 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. 3.1.2), 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[1] 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.

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



(a) (b) (c)

Figure 3.1: 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.

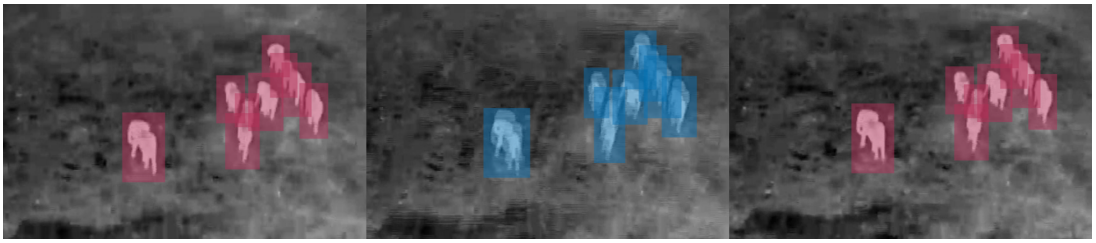


Figure 3.2: Consecutive frames from a video in the dataset showing noise. Blue colored labels are noisy labels, while red are normal animal labels.

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. 3.2 and Fig. 3.3, 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.

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. 3.1. Different statistics over the entire dataset, including distribution of object scales and densities across the train/test splits, are shown in Fig. 3.1 (a) and (b), respectively. In Fig. 3.1 (c), a scatter plot of tracking video sequences is shown with respect to the sequence length and average object density.

3.1.2 Dataset Properties

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

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. 3.4 (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[2]. 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. 3.1. There is an example of a video with high animal

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

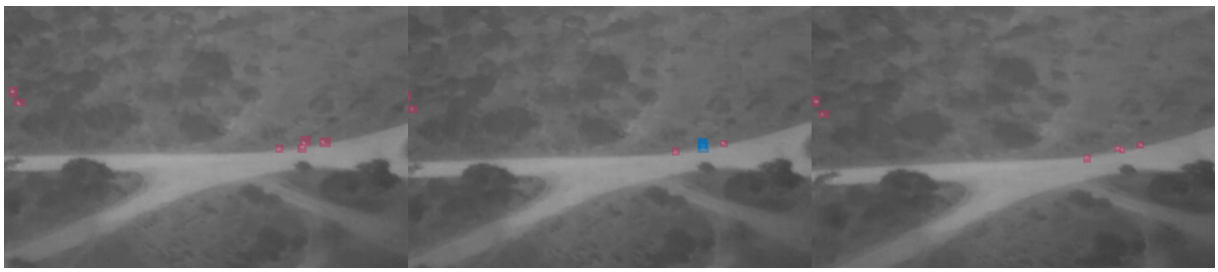


Figure 3.3: Near-consecutive frames from a video in the dataset showing occlusion. Blue colored labels are occlusion labels, while red are normal animal labels.

density in Fig. 3.4 (a).

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. 3.4 (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. 3.4 (c). We also see thermal reflections off the ground, typically near trees, e.g., in Fig. 3.4 (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 [6,61], the most common type in BIRDSAI is what we call ghosting, as shown in Fig. 3.4 (e). There are also slightly more mild versions of it, which look like horizontal "bands" in some cases. Additionally, the UAV'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. 3.4 (f). These were labeled as containing noise when possible (see Sec. 3.1.1).

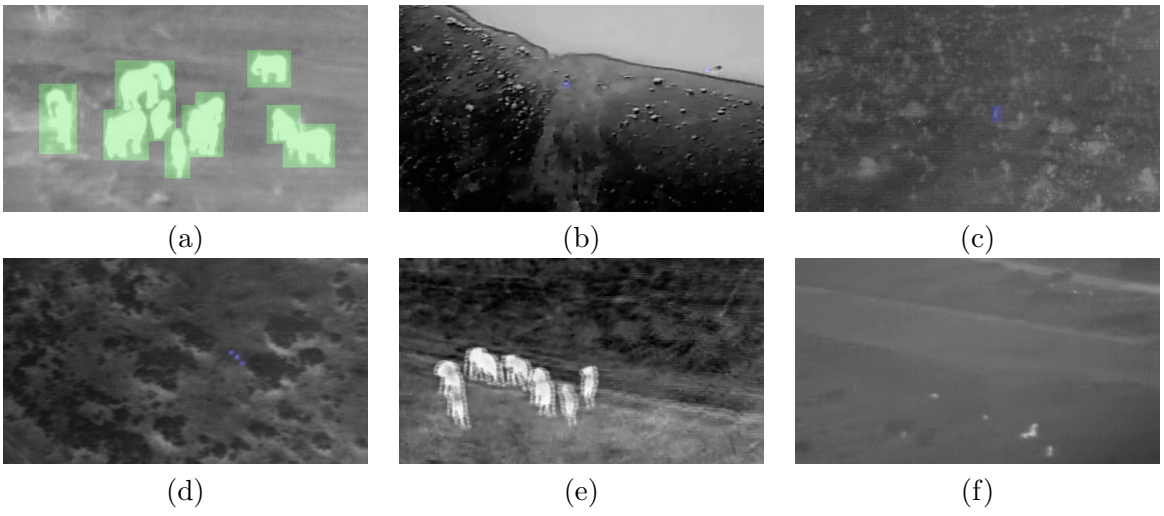


Figure 3.4: Data challenges. (a) density (b) high contrast (c) clutter (vegetation) (d) clutter (reflections) (e) ghosting (f) motion blur. Ground truth labels not shown in (e) and (f) for better visualization of effects of noise. Animals in (a), (e), (f), humans in (b), (c), (d).

Table 3.1: Details of the datasets used for training and performance evaluation. The table shows the number of cameras (#Cameras), duration of the videos, frame rate (FPS), the number of targets (#Target) captured in each dataset.

	#Cameras	Duration	FPS	#Target
NLPR-Set1	3	20 min	20	235
NLPR-Set2	3	20 min	20	255
NLPR-Set3	4	3.5 min	25	14
NLPR-Set4	5	24 min	25	49
DukeMTMC	8	1hr 25min	60	2834

3.2 Re-identification Approach

In this work, we study Re-ID networks in two ways. For Re-ID, we use the state of the art ABDNet [13][3]. The network creates feature vectors for an object, which are used to find distances between images. These distances are used to rank query images.

First, we study its performance in a Re-ID task - Given a gallery set G (template images) of a target object and a set of query images Q, identify which image in Q is that of the target, if any. We start with the network pre trained on DukeMTMC [57] dataset.

In the second analysis, we integrate this Re-ID network in a larger Camera Selection Network to study its performance in Inter Camera Tracking (ICT) and Single Camera Tracking (SCT) experiments. In SCT, we find the trajectry of a target within a single camera. In ICT, the aim is to associate trajectories with same identity across multiple cameras. For the SCT experiment, Re-ID is used to associate the bounding boxes of the objects in one frame to their corresponding bounding boxes in the next frame, thus finding the trajectory of each target. For ICT experiments, we query different cameras and search for the target in the query images. The camera to query is provided by the camera selection model, while Re-ID is used to search for the target within the query bounding boxes from the query camera.

3.2.1 Datasets

We use NLPR_MCT data set [44] and DukeMTMC and DukeRe-ID [57] dataset to evaluate the Re-ID performance independently (Re-ID) task as well as integrated with a method for camera selections in multi-camera networks.

NLPR dataset consists of four sub-datasets each having different number of cameras. NLPR-Set1 and NLPR-Set2 consists of 3 cameras from a footpath, NLPR-Set3 consists of 4 cameras of an indoor office building, and NLPR-Set4 captures the parking environment in 5 cameras. The DukeMTMC dataset contains 8 cameras of Duke University campus. Details of the dataset are in table 3.1

[3]Author's implementation `https://github.com/TAMU-VITA/ABD-Net`

3.2.2 Re-identification techniques

The ABDNet takes as input cropped images of the objects and encodes them in a feature representation. These features are then use to determine distances between two objects, lower distance indicating that the objects have the same identity. Given a gallery set G with n images and a set of query images Q with m images, we use ABDNet to find the feature vectors of each image in both G and Q, giving us feature sets F_g and F_q respectively. Then, we find pairwise euclidean distance between the feature vectors. This results in a mxn distance matrix D, where $D[i, j] = dist(F_q[i], F_g[j])$. We explore various ways of using the distances in this distance matrix to identify the target object, tuning each to give optimal performance:

- Mean Across Gallery
  We take the mean distance across the rows of the matrix to find the mean distance of each query image with all the template images, giving us D_m. Where $D_m[i] = \frac{\sum_{j=1}^{n} D[i,j]}{n}$
- Maximum Distance Threshold
  If the minimum distance in the entire distance matrix is greater than a particular threshold value, then the target is not present in the frame. We explored both a general threshold and a dataset specific thresholds
- Rank Confidence
  When there are multiple items in the frame, we define the confidence to be the ratio of second minimum and the minimum values in D_m, where a higher confidence indicates a better chance of correctly identifying the target. This confidence also has a threshold which overrides the Maximum Distance Threshold
- Adaptive Gallery Set
  Instead of initialising the template images of the target from only the first camera, we randomly select them from all the cameras in the target's trajectory.

3.2.3 Camera Selection Network

We integrate the ReID network into a Reinforcement Learning based Camera Selection network. The task of the camera selection is to predict the time and camera to look for a target based on its previous trajectory. The architecture is shown in figure 3.5. In the ICT task, the network predicts the the camera and the frame (time-point) to look for a particular target. The gallery set, or template images, of that target along with the camera and frame are given to the ReID architecture, which then compares the detected objects in the frame and returns whether or not the target exists in the frame, and which bounding box is that of the target.

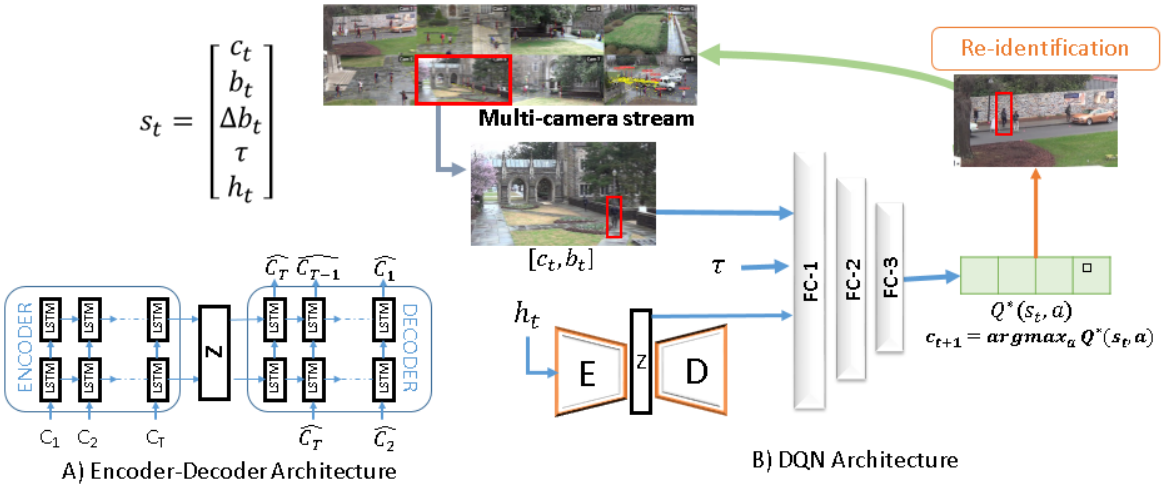


Figure 3.5: The DQN based Camera Selection Model

Chapter 4

Evaluation

4.1 Baselines for BIRDSAI

The goal of BIRDSAI is to advance image-based object detection, domain adaptive detection, and single and multi-object tracking (SOT and MOT, respectively). 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 [45], ECO [23] and AD-Net [81] algorithms as benchmarks, and we also use the MCFTS [50] 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 [47, 75]. 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 [75], 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. We show single object tracking (SOT) performance over the perfect subsequences and full sequences using the standard tracking

metrics in Fig. 4.1 and Fig. 4.2, respectively.

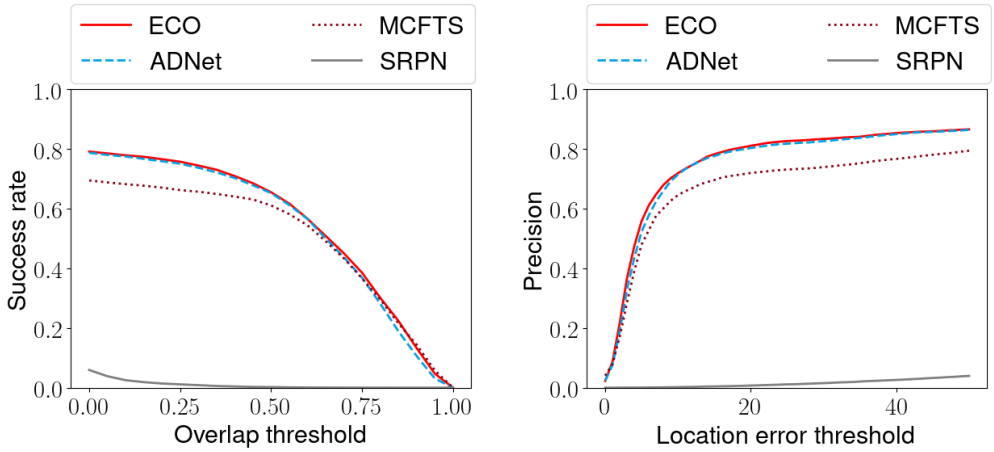


Figure 4.1: Success and precision plots for the SOT with benchmark algorithms on *perfect subsequences*.

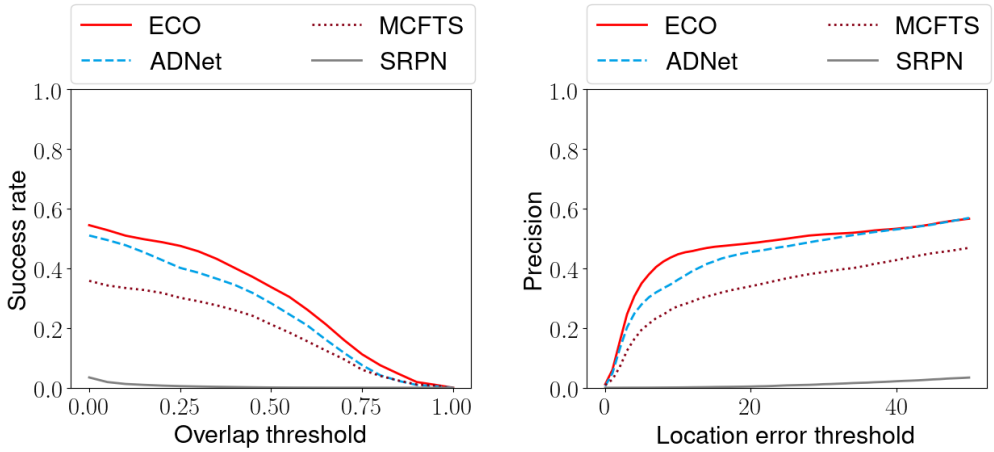


Figure 4.2: Success and precision plots for the SOT with benchmark algorithms on the entire set of *full sequences*.

For multi-object tracking we only report the IoU Tracker [5] with default thresholds, and object 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 [58], 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 4.1 for SOT and Table 4.2 for MOT benchmarking. We observe that Siamese RPN [45] 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 thermal

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 4.1: 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**
MDP Tracker (GT init.)	S	21.6	75.9	-	-
	M	54.6	84.1	-	-
	L	75.8	90.8	-	-

Table 4.2: 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.

infrared sensing modality, and the sometimes large camera motion. ECO [23] 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 [50] 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 [81] 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. 4.1 and 4.2. In most real-world scenarios, the sequences will be affected by noise, occlusions, the object leaving the frame and other such interruptions.

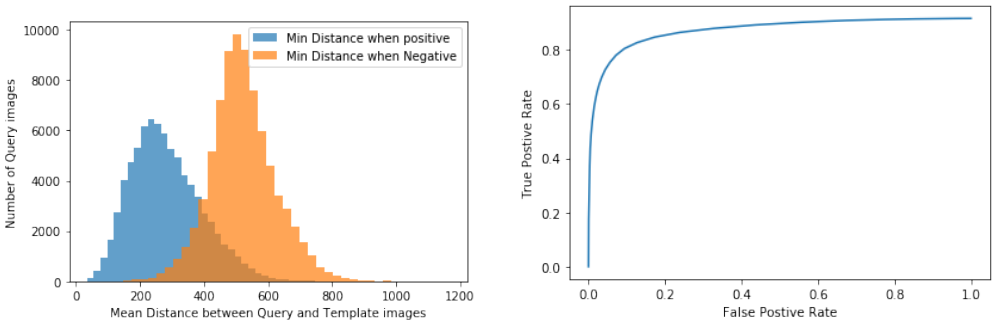
4.2 Target Re-identification

4.2.1 Re-identification Experiments

To evaluate ABDNet, we conduct person reidentification experiments on Duke-MTMC-ReID and NLPR datasets. For this analysis, we use standard Re-ID metrics - Rank1 Accuracy and mean Average Precision (mAP) [13].

Table 4.3: Re-ID results for pretrained ABDNet

	NLPR1	NLPR2	NLPR3	NLPR4	Duke
Rank1	58.62	56.82	40.00	75.86	77.59
mAP	44.93	42.84	35.71	50.39	88.33



(a) Histogram (b) ROC

Figure 4.3: Histogram of distance between query and gallery image. The ROC is calculated by using various thresholds for classifying a query image as positive and negative.

Table 4.3 shows the ReID results for ABDNet pretrained on DukeMTMC-ReID dataset.

To determine the preliminary starting value of the hyperparameters mentioned in 3.2.2, we analyze ReID performance during the transition periods of a target in the Duke dataset. Transition is the time after the target exits a camera and before it enters the next camera. For each target t we make the gallery set or template set as the first 5 images of the target's entire trajectory. We collect all frames where the target is transitioning from one camera to another. Figure 4.3 shows the histogram of number of query images with a particular distance from the gallery set. X axis shows the distance and Y axis shows the number of query images with that distance from the gallery template. Blue histogram is when query image is positive (i.e. of the same object) and Orange histogram is when query image is negative (i.e. not or the same object as the gallery set). Clearly, a threshold of 400 is a good tradeoff between less false positives and more true positives.

4.2.2 Multi Target Multi Camera Tracking Experiments

Here, we will show the tracking performance while tracking the target using our architecture explained in section 3.2.3. We will also compare the tracking performance with state-of-the-art methods on NLPR and DukeMTMC datasets using MCTA metric. For this experiment, the initial position of a target and the state representation of zero-initialized action history are used to make the initial state. The initial state is then used by the camera selection network to select a camera where the target is expected to reappear at the next time instant. Then a re-identification query is made to the selected camera to identify whether the target is present in the selected camera. If the target is present then the location of the target is updated accordingly in the state vector otherwise the time-elapse is updated. The procedure is repeated until the video

sequence ends. Please note that our method is single target multi-camera tracking approach and to make it train multiple targets, we run multiple parallel pipeline of our method starting from the initial location of the target.

Camera selections inherently improves the tracking performance as shown in the table 4.4 which shows the tracking performance of our methods and other methods on NLPR and DukeMTMC datasets. The methods in the table are separated based on how these methods resolve the camera handover (re-identifying the target). In the table, *Self* means that the methods have proposed their own approach to resolve the camera handover, GT signifies that methods use ground truth for resolving the handover, and *ReID* means that a re-identification method in [13] is used. The approach *Topology* is a baseline method where we assume that the camera topology is known and the neighboring cameras are queries to resolve the camera handover. The dashed values means that a method doesn't report those results. In experiment-1, only the inter-camera tracking (ICT) performance is evaluated. For this, detection and single camera tracking are taken from the ground truth. The camera selection decisions are taken at all times during ICT and a re-identification query is made when the model provides a valid camera input (the camera selection model provides invalid camera numbers to skip that frame). In experiment-2, only the detections are taken from ground truth. The camera selection network is used at all times both when the target is transitioning and moving in a particular camera FOV. A re-identification query is resolved accordingly.

The table shows that our camera selection method performs better on most of the datasets. On replacing the ground truth association with ReID association, we see that the performance drops which is expected. But the main observation here is that even though the performance of ABDNet ReID was poor on some of the datasets (as in Table 4.3, the performance drop is not huge and our camera selection model *still* outperforms other state-of-the art methods which use ground truth associations in Experiment 1. The performance actually improves for NLPR dataset 4! In Experiment 2, however, the drop in performance is significant after applying ReID. Although our method is not able to outperform other methods which use ground-truth association, it is still comparable and in some cases better than state-of-the-art methods using their own associations.

This shows promise in use of ReID especially in multi camera reidentification of targets. As disucssed in 3.1.2 TIR domain and UAV domain make our BIRDSAI dataset challenging, and application of ReID to develop a new tracking approach could lead to good results in tracking, target identification, species identification etc.

Table 4.4 shows the results of the To evaluate the inter-camera tracking and multi-camera tracking performance, we use commonly used Multi-Camera Tracking Accuracy (MCTA). metric [44].

Table 4.4: The table is showing average MCTA values for inter-camera tracking (ICT) and both SCT-ICT on the test set of NLPR_MCT and DukeMTMC dataset. The results are separated based on the type of association method. *Self* means a method uses its own association, *GT* represents ground truth, and *ReID* signifies that a re-identification method is used for association. We used ABDNet [13] for ReID. The dashed values means that a method doesn't report those results.

		Inter-camera tracking (ICT)					**Single-camera tracking + ICT**				
Approach	Association	Set-1	Set-2	Set-3	Set-4	Duke	Set-1	Set-2	Set-3	Set-4	Duke
[78]	Self	0.9152	0.9132	0.5163	0.7152	-	0.8831	0.8397	0.2427	0.4357	-
[15]	Self	0.7425	0.6544	0.7369	0.3945	-	0.7477	0.6561	0.2028	0.2650	-
[14]	Self	0.6617	0.5907	0.7105	0.5703	-	0.6903	0.6238	0.0848	0.1830	-
[17]	Self	0.3203	0.3456	0.1381	0.1562	-	0.8162	0.7730	0.1240	0.4637	-
[44]	Self	0.9610	0.9264	0.7889	0.7578	-	-	-	-	-	-
[70]	Self	0.835	0.703	0.742	0.385	-	0.8525	0.7370	0.4724	0.3778	-
CamSel [63]	GT	0.8210	0.7498	0.9099	0.8993	OM	0.8235	0.7503	0.9134	0.9118	OM
nSteps [64]	GT	0.9016	0.8741	0.9038	0.8074	0.8027	0.9018	0.8806	0.9058	0.7871	0.8191
Ours	GT	**0.968**	**0.963**	**0.914**	0.759	**0.902**	**0.966**	**0.961**	**0.906**	**0.776**	**0.894**
Topology	ReID	0.6405	0.3627	0.2618	0.5386	0.9784	0.5119	0.2564	0.1445	0.4426	0.5487
Ours	ReID	0.9292	0.8806	0.8426	**0.7808**	0.8855	0.7639	0.7594	0.3547	0.5258	0.7308

Chapter 5

Conclusion

We presented BIRDSAI, a challenging dataset containing aerial, TIR images of protected areas for object detection, 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.

To that end, we explore the use of ReID for multi camera re-identification and tracking. We study a state-of-the-art ReID network in a standard ReID task, multi camera transition and finally integrating it into a larger camera selection model for mutli target multi camera (MTMC) tracking experiment. We find that even though the ReID network under consideration performed poorly on the standard ReID task when compared to all the targets in the dataset, using it in real world applications showed promising results for its application in MTMC tracking.

Future work includes extending the above research to develop a new tracking approach for UAV-TIR wildlife monitoring, using approaches like ReID, Superpoint [27] etc to tackle the challenges this domain poses. 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.

Bibliography

[1] Air Shepherd. Air shepherd: The lindbergh foundation. `http://airshepherd.org`, 2019. Accessed: 2019-11-02.

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

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

[4] Bo Wu, and Nevatia, R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Oct 2005), vol. 1, pp. 90–97 Vol. 1.

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

[6] Bondi, E., Dey, D., Kapoor, A., Piavis, J., Shah, S., Fang, F., Dilkina, B., Hannaford, R., Iyer, A., Joppa, L., and Tambe, M. Airsim-w: A simulation environment for wildlife conservation with uavs. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (2018), COMPASS '18, pp. 40:1–40:12.

[7] Bondi, E., Fang, F., Kar, D., Noronha, V., Dmello, D., Tambe, M., Iyer, A., and Hannaford, R. Viola: Video labeling application for security domains. In *Proceedings of the 8th Annual Conference on Decision Theory and Game Theory for Security (GameSec)* (2017).

[8] Bondi, E., Jain, R., Aggrawal, P., Anand, S., Hannaford, R., Kapoor, A., Piavis, J., Shah, S., Joppa, L., Dilkina, B., et al. Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. In *The IEEE Winter Conference on Applications of Computer Vision* (2020), pp. 1747–1756.

[9] Boon, M., P. Drijfhout, A., and Tesfamichael, S. Comparison of a fixed-wing and multi-rotor uav for environmental mapping applications: A case study. *ISPRS - Interna-*

tional Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W6 (08 2017), 47–54.

[10] Chari, V., Lacoste-Julien, S., Laptev, I., and Sivic, J. On pairwise costs for network flow multi-object tracking. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 5537–5545.

[11] Cheema, G. S., and Anand, S. Automatic Detection and Recognition of Individuals in Patterned Species. In *ECML PKDD* (2017).

[12] Chen, K. W., Lai, C. C., Lee, P. J., Chen, C. S., and Hung, Y. P. Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras. *IEEE Transactions on Multimedia 13*, 4 (Aug 2011), 625–638.

[13] Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., and Wang, Z. Abd-net: Attentive but diverse person re-identification, 2019.

[14] Chen, W., Cao, L., Chen, X., and Huang, K. A novel solution for multi-camera object tracking. In *2014 IEEE International Conference on Image Processing (ICIP)* (Oct 2014), pp. 2329–2333.

[15] Chen, W., Chen, X., and Huang, K. Multi-Camera Object Tracking (MCT) Challenge. `http://http://mct.idealtest.org/Datasets.html/`, 2014.

[16] Chen, X., An, L., and Bhanu, B. Multitarget tracking in nonoverlapping cameras using a reference set. *IEEE Sensors Journal 15*, 5 (May 2015), 2692–2704.

[17] Chen, X., and Bhanu, B. Integrating social grouping for multitarget tracking across cameras in a crf model. *IEEE Transactions on Circuits and Systems for Video Technology 27*, 11 (Nov 2017), 2382–2394.

[18] Christiansen, P., Steen, K., Jørgensen, R., and Karstoft, H. Automated detection and recognition of wildlife using thermal cameras. *Sensors 14*, 8 (2014), 13778–13793.

[19] Collins, R. T. Multitarget data association with higher-order motion models. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (June 2012), pp. 1744–1751.

[20] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

[21] Crall, J., Stewart, C., Berger-Wolf, T., Rubenstein, D., and Sundaresan, S. Hotspotter – patterned species instance recognition. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on* (2013), pp. 230–237.

[22] Daliyot, S., and Netanyahu, N. S. A framework for inter-camera association of multi-target trajectories by invariant target models. In *Computer Vision - ACCV 2012 Workshops* (Berlin, Heidelberg, 2013), J.-I. Park and J. Kim, Eds., Springer Berlin Heidelberg, pp. 372–386.

[23] Danelljan, M., Bhat, G., Shahbaz Khan, F., and Felsberg, M. Eco: Efficient convolution operators for tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[24] Das, A., Chakraborty, A., and Roy-Chowdhury, A. K. Consistent re-identification in a camera network. In *Computer Vision – ECCV 2014* (Cham, 2014), D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Springer International Publishing, pp. 330–345.

[25] Dehghan, A., Assari, S. M., and Shah, M. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015), pp. 4091–4099.

[26] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR* (2009), IEEE, pp. 248–255.

[27] DeTone, D., Malisiewicz, T., and Rabinovich, A. Superpoint: Self-supervised interest point detection and description. *CoRR abs/1712.07629* (2017).

[28] Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., and Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In *The European Conference on Computer Vision (ECCV)* (September 2018).

[29] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision 88*, 2 (June 2010), 303–338.

[30] Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 2 (Feb 2008), 267–282.

[31] Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 2 (Feb 2008), 267–282.

[32] Freytag, A., Rodner, E., Simon, M., Loos, A., Kühl, H. S., and Denzler, J. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition* (2016), Springer, pp. 51–63.

[33] Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2012).

[34] Graves, H. B., Bellis, E. D., and Knuth, W. M. Censusing white-tailed deer by airborne thermal infrared imagery. *The Journal of Wildlife Management 36*, 3 (1972), 875–884.

[35] Hwang, S., Park, J., Kim, N., Choi, Y., and Kweon, I. S. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1037–1045.

[36] Javed, O., Shafique, K., Rasheed, Z., and Shah, M. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Comput. Vis. Image Underst. 109*, 2 (Feb. 2008), 146–162.

[37] Jiang, N., Bai, S., Xu, Y., Xing, C., Zhou, Z., and Wu, W. Online inter-camera trajectory association exploiting person re-identification and camera topology. In *Proceedings of the 26th ACM International Conference on Multimedia* (New York, NY, USA, 2018), MM '18, Association for Computing Machinery, p. 1457–1465.

[38] Kamminga, J., Ayele, E., Meratnia, N., and Havinga, P. Poaching detection technologies—a survey. *Sensors 18*, 5 (2018).

[39] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. Large-scale video classification with convolutional neural networks. In *CVPR* (2014).

[40] Kellenberger, B., Marcos, D., and Tuia, D. Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment 216* (2018), 139 – 153.

[41] Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernández, G., Vojir, T., Hager, G., Nebehay, G., and Pflugfelder, R. The visual object tracking vot2015 challenge results. In *ICCV Workshops* (2015), pp. 1–23.

[42] Kumar, S., and Singh, S. K. Visual animal biometrics: survey. *IET Biometrics 6*, 3 (2017), 139–156.

[43] Kuo, C.-H., Huang, C., and Nevatia, R. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *Proceedings of the 11th European Conference on Computer Vision Part I* (Berlin, Heidelberg, 2010), ECCV2010, Springer-Verlag, pp. 383–396.

[44] Lee, Y., Tang, Z., Hwang, J., and Y. Online-learning-based human tracking across non-overlapping cameras. *IEEE Transactions on Circuits and Systems for Video Technology 28*, 10 (Oct 2018), 2870–2883.

[45] Li, B., Yan, J., Wu, W., Zhu, Z., and Hu, X. High performance visual tracking with siamese region proposal network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[46] Li, C., Liang, X., Lu, Y., Zhao, N., and Tang, J. RGB-T object tracking: Benchmark and baseline. *CoRR abs/1805.08982* (2018).

[47] Li, S., and Yeung, D.-Y. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *AAAI* (2017).

[48] Li Zhang, Yuan Li, and Nevatia, R. Global data association for multi-object tracking using network flows. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (June 2008), pp. 1–8.

[49] Liu, Q., and He, Z. PTB-TIR: A thermal infrared pedestrian tracking benchmark. *CoRR abs/1801.05944* (2018).

[50] Liu, Q., Lu, X., He, Z., Zhang, C., and Chen, W.-S. Deep convolutional neural networks for thermal infrared object tracking. *Knowledge-Based Systems 134* (2017), 189 – 198.

[51] Lyu, S., Chang, M.-C., Du, D., Wen, L., Qi, H., Li, Y., Wei, Y., Ke, L., Hu, T., Del Coco, M., et al. 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* (2017), IEEE, pp. 1–7.

[52] Ma, Y., Wu, X., Yu, G., Xu, Y., and Wang, Y. Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery. *Sensors 16*, 4 (2016), 446.

[53] Makris, D., Ellis, T., and Black, J. Bridging the gaps between cameras. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* (June 2004), vol. 2, pp. II–205–II–210 Vol.2.

[54] Mueller, M., Smith, N., and Ghanem, B. A benchmark and simulator for uav tracking. In *Proc. of the European Conference on Computer Vision (ECCV)* (2016).

[55] Olivares-Mendez, M. A., Bissyandé, T. F., Somasundar, K., Klein, J., Voos, H., and Le Traon, Y. The noah project: Giving a chance to threatened species in africa with uavs. In *e-Infrastructure and e-Services for Developing Countries* (2014), T. F. Bissyandé and G. van Stam, Eds., pp. 198–208.

[56] Portmann, J., Lynen, S., Chli, M., and Siegwart, R. People detection and tracking from aerial thermal views. In *2014 IEEE International Conference on Robotics and Automation (ICRA)* (2014), pp. 1794–1800.

[57] Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking* (2016).

[58] Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking* (2016).

[59] Ristani, E., and Tomasi, C. Tracking multiple people online and in real time. In *Computer Vision – ACCV 2014* (Cham, 2015), D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds., Springer International Publishing, pp. 444–459.

[60] Ristani, E., and Tomasi, C. Features for multi-target multi-camera tracking and re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[61] Schott, J. R. *Remote sensing: the image chain approach*. Oxford University Press on Demand, 2007.

[62] Shafique, and Shah. A non-iterative greedy algorithm for multi-frame point correspondence. In *Proceedings Ninth IEEE International Conference on Computer Vision* (Oct 2003), pp. 110–115 vol.1.

[63] Sharma, A., Anand, S., and Kaul, S. K. Reinforcement learning based querying in camera networks for efficient target tracking. In *Proceedings of International Conference on Automated Planning and Scheduling (ICAPS), 2019* (2019).

[64] Sharma, A., Anand, S., and Kaul, S. K. Intelligent querying for target tracking in camera networks using deep q-learning with n-step bootstrapping, 2020.

[65] Shu, G., Dehghan, A., Oreifej, O., Hand, E., and Shah, M. Part-based multiple-person tracking with partial occlusion handling. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (June 2012), pp. 1815–1821.

[66] Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., and Packer, C. Data from: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna, 2015.

[67] Tang, S., Andriluka, M., Andres, B., and Schiele, B. Multiple people tracking by lifted multicut and person re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), pp. 3701–3710.

[68] Tesfaye, Y. T., Zemene, E., Prati, A., Pelillo, M., and Shah, M. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *CoRR abs/1706.06196* (2017).

[69] UNODC. World wildlife crime report: Trafficking in protected species, 2016.

[70] W. Chen, L. C., Chen, X., Huang, K., Huang, K., and Huang, K. An equalized global graph model-based approach for multicamera object tracking. *IEEE Transactions on Circuits and Systems for Video Technology 27*, 11 (Nov 2017), 2367–2381.

[71] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[72] Wan, J., and Liu Li. Distributed optimization for global data association in non-overlapping camera networks. In *2013 Seventh International Conference on Distributed Smart Cameras (ICDSC)* (Oct 2013), pp. 1–7.

[73] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.

[74] Witham, C. L. Automated face recognition of rhesus macaques. *Journal of neuroscience methods* (2017).

[75] Wu, Y., Lim, J., and Yang, M.-H. Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell. 37*, 9 (2015), 1834–1848.

[76] Wu, Z., Fuller, N., Theriault, D., and Betke, M. A thermal infrared video benchmark for visual analysis. In *2014 IEEE CVPR Workshop on Perception Beyond the Visible Spectrum* (2014), pp. 201–208.

[77] Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[78] Y. Cai, G. M. Exploring context information for inter-camera multiple target tracking. In *IEEE Winter Conference on Applications of Computer Vision* (March 2014), pp. 761–768.

[79] Yahyanejad, S., and Rinner, B. 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* (2015), 189–202.

[80] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. H. Deep learning for person re-identification: A survey and outlook, 2020.

[81] Yun, S., Choi, J., Yoo, Y., Yun, K., and Young Choi, J. Action-decision networks for visual tracking with deep reinforcement learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[82] Zhang, S., Staudt, E., Faltemier, T., and Roy-Chowdhury, A. K. A camera network tracking (camnet) dataset and performance baseline. In *2015 IEEE Winter Conference on Applications of Computer Vision* (Jan 2015), pp. 365–372.

[83] Zhao, B., Feng, J., Wu, X., and Yan, S. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing 14*, 2 (2017), 119–135.