# Supplementary Material – BIRDSAI: A Dataset for Detection and Tracking in Aerial Thermal Infrared Videos

# 1. Dataset Additional Details

# **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. 1 and Fig. 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.

## 1.2. Simulated Data

We added a lion to the simulation. We used 38  $^{\circ}$ C (311 K) for temperature in summer, 39  $^{\circ}$ C (310 K) for temperature in winter [8], and 0.98 for emissivity [7]. 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,



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



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

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.

## 2. Additional Experiments

# 2.1. Detections

Table 1 contains the results for two of the detection models on the proposed dataset, BIRDSAI, with ResNet [4] 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 2 (extension of Table 3 in the main paper, including the same Syn  $\rightarrow$  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 [2], 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 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.

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 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
$Syn \rightarrow Real$	0.443	0.459	0.309	0.313
Syn $\rightarrow$ Real (50% Sup. DA)	0.466	0.474	0.384	0.398
Syn $\rightarrow$ Real (100% Sup. DA)	0.522	0.518	0.448	0.472

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

#### 2.1.1 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 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 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.

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

### 2.2. Tracking

## 2.2.1 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 4 across different tracking algorithms - ADNet<sup>1</sup>, ECO<sup>2</sup>, Siamese RPN<sup>3</sup> and MCFTS<sup>4</sup> - for ease. We show single object tracking (SOT) performance over the perfect subsequences and full sequences using the standard tracking metrics in Fig. 3 and Fig. 4, respectively.

We observe that Siamese RPN [5] 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 infrared sensing modality, and the sometimes large camera motion. ECO [3] 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 [6] 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 [10] 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. 3 and 4. In most real-world scenarios, the sequences will be affected by noise, occlusions, the object leaving the frame and other such interruptions.

#### 2.2.2 Multi Object Tracking

Table 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 [9] underperforms the IoU tracker, when the latter is provided with ground truth detections.

## 2.3. Cross-Dataset Evaluation

We also provide results trained using the LTIR dataset [1], 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 6. Based on these results, we conclude that the BIRDSAI dataset is substantially different than [1]. Moreover, based on the results in the previous sections, we can also conclude that it is sufficiently challenging by itself.

https://github.com/hellbell/ADNet

<sup>&</sup>lt;sup>2</sup>https://github.com/martin-danelljan/ECO

<sup>&</sup>lt;sup>3</sup>https://github.com/songdejia/Siamese-RPN-pytorch

<sup>&</sup>lt;sup>4</sup>https://github.com/QiaoLiuHit/MCFTS



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



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

Method	Perfect Sub	sequences	Full Sequence	
Methou	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: Single Object Tracking Evaluation. Precision is at 20 pixels. "Perfect subsequences" excludes noisy/occluded frames, while "Full sequence" includes them.

# References

- [1] A. Berg, J. Ahlberg, and M. Felsberg. A thermal object tracking benchmark. In 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–6, 2015. 3
- [2] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In Computer Vision and Pattern Recognition (CVPR), 2018. 2
- [3] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In The

Method	Object Size	MOTA	MOTP
IoII Trookor	S	61.6	100.0
(CT dat)	М	91.3	98.9
(Of det.)	L	80.6	100.0
MDD Troolcor	S	21.6	75.9
(GT init)	М	54.6	84.1
(01 mit.)	L	75.8	90.8

Table 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 6: Cross-Dataset Detection p	erformance evaluation	using the mAP metric
------------------------------------	-----------------------	----------------------

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017. 3

- [4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2
- [5] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3
- [6] Q. Liu, X. Lu, Z. He, C. Zhang, and W.-S. Chen. Deep convolutional neural networks for thermal infrared object tracking. *Knowledge-Based Systems*, 134:189 – 198, 2017. 3
- [7] D. J. Mccafferty. The value of infrared thermography for research on mammals: previous applications and future directions. *Mammal Review*, 37(3):207–223, 2007. 1
- [8] P. Trethowan, A. Fuller, A. Haw, T. Hart, A. Markham, A. Loveridge, R. Hetem, B. du Preez, and D. W. Macdonald. 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, 2017. 1
- [9] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *Interna*tional Conference on Computer Vision (ICCV), 2015. 3
- [10] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3