Targeted Image Transformation for Improving Robustness in Long Range Aircraft Detection

Rebecca Martin[1], Clement Fung[2], Nikhil Keetha[1], Lujo Bauer[2], Sebastian Scherer[1]

***Abstract*— In the field of aviation, the Detect and Avoid (DAA) problem deals with incorporating collision avoidance capabilities into current autopilot navigation systems. As an application of the Small Object Detection (SOD) problem, DAA presents the difficulties of a low signal-to-noise ratio and far range detection. Visual DAA is also susceptible to changing weather and lighting conditions at deployment. While current literature has presented many solutions for this, prior work has yet to study the robustness of the learning-based models for DAA. In this work, we show that standard techniques for improving robustness for object detection do not produce the desired results for DAA given the SOD constraints. We present targeted transformations, a zero-shot technique that can significantly improve robustness with minimal impact on accuracy. We demonstrate how to construct these transformations and evaluate our method on the current SOTA model for DAA, showing a 53.6% increase in recall. This makes our pipeline more robust to changes in lighting and environmental factors, and better able to detect potential threats. In the future, we hope to automate the transformation selection process, making it easier to adopt in different use cases.**

I. Introduction

As autonomous cars and planes are quickly becoming a reality, it is critical to be able to assess the robustness of their perception systems. Risk detection systems are especially important for autonomous planes, as they operate in a much larger and less structured environment. In the case of camera-based detection systems, seeing these potential risks from far distances falls in the category of Small Object Detection (SOD). Small Object Detection, as a sub-field of generic object detection, concentrates on detecting objects of small size and is of great significance in various scenarios such as surveillance, drone scene analysis, pedestrian detection, traffic sign detection in autonomous driving, etc.,where accurate detection is still needed from as far as 100m away.

Within the field of aviation, the Detect and Avoid (DAA) problem deals with incorporating collision avoidance capabilities into current autopilot navigation systems. Detect and Avoid is defined by the Federal Aviation Administration as "the capability of an aircraft to remain well clear from and avoid collisions with other airborne traffic." [34] The deployment constraints of Visual DAA, such as the low signal-to-noise ratio and far range of the detection inputs, make the detection phase of this collision avoidance problem a clear application of SOD.

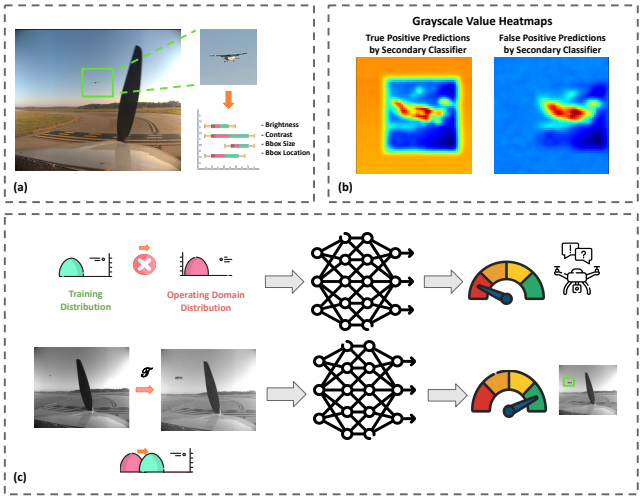

Fig. 1: We explore methods beyond benchmarking to analyze the robustness of perception methods for Detect and Avoid: (a) We examine how properties of input image data can have a downstream effect on model performance. (b) Controlled experiments can lead to the discovery of biases or shortcuts learned by a model. For example, an experiment examining the average grayscale values for different predictions made by a false positive filter (in our case, a secondary classifier in the AirTrack [12] model) shows that any dark object on a lighter background is determined as a true positive (warmer heatmap indicates lighter grayscale values). (c) Transformations in the image space can help align runtime data to training distribution, leading to more robust performance.

While current literature contains many methods for approaching this problem and solving it with high precision, prior work has yet to study the robustness of these solutions for both DAA and SOD in general. Most prior works focused on robustness consider image detection and classification tasks with a high signal-to-noise ratio [20], [7], and our DAA use case involves two unique challenges that are unaddressed in this prior work. First, the DAA task involves object detection, in which the model predicts a bounding box rather than a ground-truth class. This eliminates prior techniques that require explicit classification labels to separate points in latent space [15]. Second, the DAA task involves a very low signal-to-noise ratio: on average, a ground-truth bounding box makes up a very small percentage of the overall image ($\sim 1\%$). Since there is such a small relative amount of signal in the input image, latent-space techniques [27], [15], [21], [11] fail to capture the fine-grained details of an image when

[1]Rebecca Martin, Nikhil Keetha, and Sebastian Scherer are with Carnegie Mellon University, School of Computer Science, Robotics Institute, 5000 Forbes Ave, Pittsburgh, PA, USA. (rebecca2, nkeetha, basti)@andrew.cmu.edu)

[2]Clement Fung and Lujo Bauer are with Carnegie Mellon University, School of Computer Science, Software and Societal Systems Department, 5000 Forbes Ave, Pittsburgh, PA, USA. (clementf, lbauer)@andrew.cmu.edu)

separating images with and without ground-truth objects.

In this work, we instead explore techniques that can work well in long-range, low-signal settings and can be applied to the DAA task. We show that standard techniques for improving robustness for object detection do not produce the desired results for DAA given the SOD constraints. We present targeted transformations, a zero-shot technique that can significantly improve robustness with minimal impact on accuracy. We demonstrate how to construct these transformations and evaluate our method on the current SOTA model for DAA, showing significantly improved recall. Figure 1 provides an overview of our method: we first analyze image properties to identify model biases in our DAA framework, and then, we design and perform targeted transformations on images to improve robustness at runtime.

The rest of this paper is structured as follows: Section II presents the current literature on robustness for perception systems. Section III explains the DAA problem in detail, including the current state-of-the-art system and relevant datasets. Section IV details our method of targeted image transformation. Section V presents our experimental results and ablations. Section VI gives conclusions and future work.

II. Related Work

While the choice of a machine learning model's architecture is important for performance, its robustness and generalization capabilities largely depend on the data used to train, validate, and test it. Training datasets should model the deployment conditions as accurately as possible, but cannot exhaustively cover every scenario. To overcome this, robustness techniques focus on bridging this gap between training and real-world environments.

The noise (i.e., difference) between training and testing distributions can be classified as corruption or adversarial [45]. Corruption noise models inputs that randomly differ from the training distribution at the distribution level; one example is when the environment and weather conditions at deployment are different than those modeled by the training data and is also referred to as distribution shift. In contrast, adversarial inputs are individual inputs that do not match the training distribution; they can be rare cases belonging to the long-tailed portion of the distribution or hand-crafted inputs designed to mislead the model [28].

In this work, we focus on corruption noise, which can be addressed by domain adaptation techniques. These techniques help models generalize to new domains unseen in the training data without incurring the overhead cost of additional training. The most popular domain adaptation methods rely on fine-tuning with limited to no labeled data in the target domain. Semi-Supervised Domain Adaptation [5], [6], [35], [42] and Few-Shot Supervised Adaptation [25], [36], [43], [44] have been studied for image classification and segmentation and can be done with limited labeled data [43]. However, the number of labeled samples needed scales with the domain gap, making these methods infeasible for larger domain gaps [8]. Unsupervised Domain Adaptation methods [41] either rely on self-training with pseudo-labels [16], [19], [39], [46], [18] or adversarial training [13], [37], [38]. Other domain adaptation methods have been proposed in a variety of contexts [27], [15], [21], [11], and often rely on latent-space projections to explore new, unseen domains. Finally, another popular domain adaptation approach uses data augmentation with input transformations. These can be applied in a zero-shot manner [22] or can be coupled with a fine-tuning approach [29].

In this work, we focus on corruption noise caused by environmental changes such as weather and lighting conditions, and propose a zero-shot image transformation technique for dataset augmentation.

III. Detect and Avoid

A. Background

Detect and Avoid (DAA) is a crucial capability that aircraft require to avoid collisions with other airborne objects. In a push for DAA standardization, ASTM [1] published F3442/F3442M-20, a set of performance requirements which define safe DAA operations for UAS with a maximum dimension less than or equal to 25ft and operating at airspeeds below 100kts, without defining a specific DAA architecture. According to this standard, for safe DAA operations, the probability of track must be greater than 95% and the range estimation error of the intruder must be within 15%, along with an upper bound on the angular rate error.

Cameras and computer vision have become more popular for building DAA systems, given their passive nature and small form factor [24]. In this context, several techniques have been explored for visual DAA following a common pipeline: (a) estimation & subtraction of ego-motion, (b) dynamic object detection, (c) temporal filtering & tracking.

Traditional approaches to visual DAA have looked at employing classical computer vision techniques in the following ways: (a) optical flow-based methods [23] or image registration [30], [33] for ego-motion estimation & subtraction, (b) morphological operations to separate the background & foreground enabling object identification [4], [9], [17], [31], and (c) track-before-detect [10] and filtering methods (such as Kalman [26] and Viterbi-based) for temporal tracking and filtering of detections [4], [17].

More recently, given the advent of deep learning, neural-network-based methods have shown impressive performance for airborne object detection and tracking. One such method, Dogfight [3], leverages a two-stage spatio-temporal segmentation approach to detect drones from videos. Further building on this, TransVisDrone [32] proposes spatio-temporal transformers for aerial drone detection. Besides drone detection, approaches such as AirTrack [12] use convolutional neural networks (CNNs) for detecting various types of airborne objects.

B. Datasets

The largest publicly available dataset for Airborne Object Tracking (AOT) was introduced in 2021 by Amazon [2]. It consists of approximately 5000 sets of flight sequences, each lasting 120 seconds and captured at a rate of 10Hz, resulting

in a cumulative flight data duration of about 164 hours. This dataset has over 3.3 million annotated image frames featuring airborne objects. These images are characterized by a resolution of 2448×2048 and are rendered in 8-bit grayscale. The size of the labeled objects varies, ranging from occupying 4 to 1000 square pixels. This dataset additionally encompasses different atmospheric and visibility conditions. Specifically, around 69% of the sequences enjoy optimal visibility, 26% exhibit medium visibility, and 5% depict poor visibility.

AirTrack [12] further introduces a real-world dataset, named TartanX6C, where a multi-camera payload is used to capture data on board a Cessna 172 performing general flight & an unmanned aerial vehicle (UAV) flying towards a helicopter in a controlled setting. The dataset consists of 18 sequences where the ego aerial vehicle (ownship) is either a helicopter, UAV, or a general aviation plane (i.e., Cessna). Similar to the AOT dataset, the images are of size 2448×2048 and are rendered as 8-bit grayscale images.

In this work, we use the AOT [2] dataset as training data and the TartanX6C [12] dataset as our test set.

C. AirTrack

In our study, we use the state-of-the-art, AirTrack [12], as our model for analysis. The AirTrack system's overall design consists of four sequential modules: Frame Alignment, Detection and Tracking, Secondary Classification, and Intruder State Update. The inputs for the system are two consecutive grayscale image frames, and the system outputs a list of tracked objects with various attributes. The description of the AirTrack modules is as follows:

Frame Alignment: This module aligns successive frames to distinguish between foreground objects and camera ego-motion. It predicts optical flow between frames and confidence in the predicted flow. It operates on input frames, cropping them to focus on high-texture regions. A ResNet-34 architecture with two prediction heads is used. Training involves creating augmented input tuples and minimizing a loss between predicted and Lucas-Kanade optical flows.

Detection: The detection module consists of two cascaded parts. The primary module takes two full-resolution frames, while the secondary module processes cropped regions around top detector outputs from the primary. The cascaded network outputs maps for the center heatmap, bounding box size, offsets, track offsets, and log-scale object distance. HRNet-W32 is used as the primary detector, while HRNet-W48 is used as the crop detector. Given the marginal benefits of the crop detector, we only consider the primary detector for our study.

Tracking: The tracking approach builds on offset tracking vectors. It associates current detections with existing tracks based on predicted track offsets. If adjusted centers match within a threshold, the detection is associated with a track; otherwise, a new track is created.

Secondary Classifier: A ResNet-18 module serves as a binary classifier for false-positive rejection. It takes cropped regions around detector bounding boxes as input and aims to improve precision by rejecting false positives. This module is trained using a focal loss, and training data is mined using the primary detector and random image cropping.

Overall, the AirTrack system's sequential modules work together to align frames, detect and track objects, and reject false positives. We use the AirTrack variant trained on the entire AOT dataset. Furthermore, we use the TartanX6C dataset as our analysis benchmark.

IV. Targeted Transformations

In this section, we show that standard techniques for improving object detection robustness do not produce the desired results in the small object detection problem. We then present targeted transformations, a zero-shot technique that can significantly improve robustness with minimal impact on accuracy, and demonstrate how to construct these transformations in the DAA setting on the AirTrack model.

A. Performance Analysis

At first glance, the AirTrack pipeline works very well, producing a high accuracy and satisfying the ASTM F3442 standard [1] for some classes of aircraft (i.e., over 95% chance of successfully tracking an aircraft up to 700m range and a range estimation error less than 15% for an aircraft up to 1.5km away). We show the range estimation error and tracking probability of AirTrack in Figures 2 and 3.

However, when certain properties of the input image are varied, the performance of the DAA model may drop considerably, indicating a spurious correlation between these image properties and the model output. We define a set of measurements for each input image; these measurements capture properties of the image that can reasonably change based off of different environmental and lighting conditions in deployment conditions. Table I lists all the image properties considered as well as how they were measured and what physical change they are modeling. Here, "global" and "local" refer to the pixel area of the image; Global properties were measured across the entire image frame, and local properties were measured within the bounding box of the object being detected and tracked.

Measurements like an image's brightness or contrast are measured at different scales. For each frame, we measure

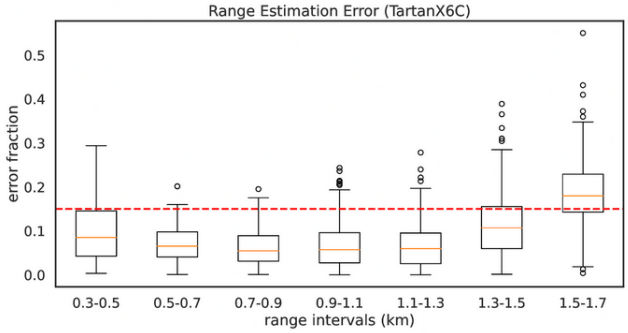

Fig. 2: Range estimation error of AirTrack pipeline, from [12]. The red line denotes the maximum average error that the system can have under the ASTM guidelines. AirTrack satisfies this standard up to a distance of 1.5km.

TABLE I: Image properties considered for transformations. For each property, we list the method and frame region used to compute it, as well as the physical phenomena that makes it relevant to our application.

Property	Global	Local	Computation	Modelled Phenomena
Brightness	X	X	Grayscale intensity mean	Sensor exposure change
Contrast	X	X	Grayscale intensity variance	Haze, overfitting to keypoints
White noise	X		Peak Signal-to-Noise Ratio	Sensor noise
Box size		X	Area of box	Planar pose and viewpoint changes
Box position		X	Relative X, Y of box center	Relative viewpoint

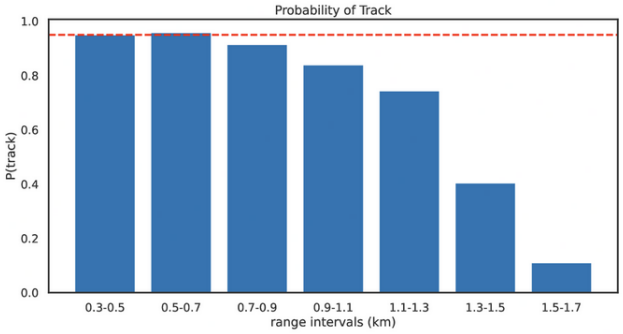

Fig. 3: Probability of track error across different ranges, from [12]. The red line denotes the minimum track probability that the system can have under the ASTM guidelines. AirTrack satisfies this standard up to a distance of 700m.

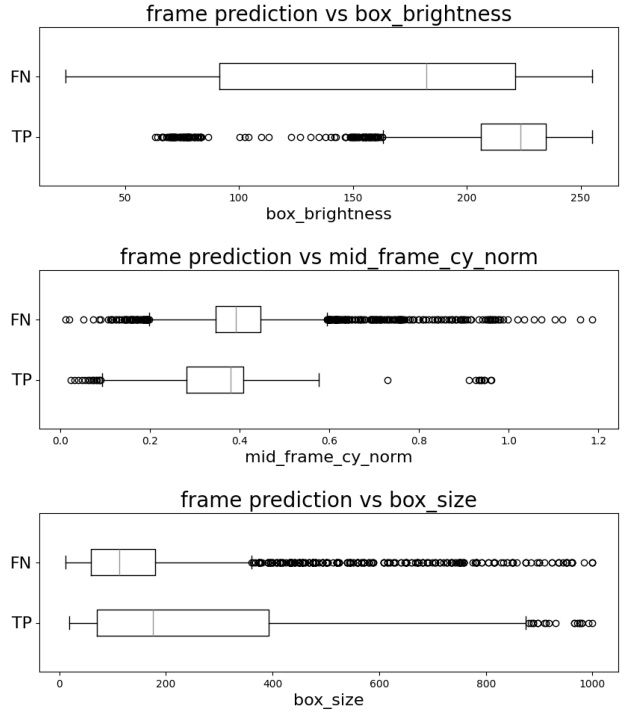

Fig. 4: Box-and-whisker plot plotting the correlation between AOT [2] image properties and AirTrack [12] model performance. There is a clear difference between the distributions of the false negative (FN) and true positive (TP) detections. For brightness and y-axis position, the distribution range is around half for true positives as for false negatives. This indicates that these properties are potential features that impact the model's detection output.

image properties in the areas local to the ground-truth object, based on the center of the ground-truth labelled bounding box. For example, we measure the *local* brightness of an object by computing the average pixel intensity in a 32x32 box centered on the ground-truth bounding box. Table I lists each image property, its measurement scope, how it is computed over each frame in our dataset, and what physical phenomena is being modelled by it.

Given a full set of measurements for each frame, we next correlate the performance of AirTrack across these properties, without using any transformations or robustness techniques. Three of these measurements provide insight into existing biases in the DAA model—we find that AirTrack performs worse when: (i) the brightness is lower, (ii) the bounding box is small, and (iii) the bounding box position is lower on the y-axis. (ii) and (iii) are both natural byproducts of the problem setting: (ii) is intuitive as smaller objects will naturally be harder to detect, and (iii) is because the sky provides a cleaner background for detection than the ground along with all objects below the horizon line. The image brightness, however, is an artifact of the dataset, as it was recorded on sunny, good weather days. This will not always be the case at deployment time, so the model should be robust to these variations in lighting.

Figure 4 shows the box plots of these measurements for each frame, with each observation categorized as a true positive or false negative based on model performance. There is a clear difference between the distributions of the false negative (FN) and true positive (TP) detections for these measurements, indicating a correlation between the image properties and whether the model detected the object. The range of brightness measurements for the true positive detections is much smaller than that of the false negatives. From the y-axis position measurements, we see that nearly all objects in the lower half of the frame were not detected, and thus a false negative. For box size, the distribution of false negatives is heavily skewed towards smaller objects. These results show that, although the overall performance of AirTrack is strong, there still exists conditions where the system is more likely to fail.

B. *Image Transformations for Adaptation*

In the previous section, we found that local image properties such as brightness and bounding box size affected the performance of AirTrack. Given that position and bounding box size were specific to the object, not the frame, and

Fig. 5: Example best transformation after image property analysis. The left image is a sample frame taken from the TartanX6C [12] dataset without any transformation. The right image is the same frame after transformation. The frame brightness has been increased with a 1.5 gamma correction and the brightness of the area around the candidate objects have been decreased with a 0.5 gamma correction.

therefore required knowledge of ground truth, we focused on brightness for our experiments. To improve robustness in the cases where AirTrack performs worst, we propose to use a gamma correction [14], which can adjust the brightness of an image while minimizing the information loss. To adjust the brightness of the frames, we applied a gamma correction instead of additively changing the brightness, so as not to lose information in the transformation process. The gamma property of an image defines "the relationship between a pixel's numerical value and its actual luminance" [14] and is typically utilized in image compression, as it is able to store pixel intensities more efficiently. From Figure 4, we can see that the frames of the true positive predictions are much brighter than those of the false positive predictions, so we transform the entire frame to increase the brightness.

As presented in Section II, state-of-the-art image augmentations for improving robustness apply transformations uniformly across a frame. While this works well for generic object detection tasks, it does not provide any benefit for Small Object Detection, which we provide empirical comparisons for in the next section. In fact, depending on the environmental conditions, the transformation can decrease detection performance. To mitigate this, we propose Targeted transformations, which consists of applying different transformations in the global frame and in the local area around the object that needs more attention. In our application, we use a gamma correction to *increase* the brightness of the entire frame and apply a separate gamma correction within the local area around the object to *decrease* the brightness in this region, using the contrast to increase model attention in that area. Figure 5 shows an example of this transformation. Detailed ablation studies of transformation composition and selection of candidate bounding boxes for the objects are presented in the next section.

V. Experimental Analysis

A. Experimental Setup

For our experiments, we used the AirTrack pipeline, trained on the Airborne Object Tracking [2] dataset as presented in [12], as our base model for all the robustness transformations and ablations presented in the following experiments. We applied our data augmentations to the TartanX6C [12] data, our chosen test dataset, made up of real world data collected from a camera on a small airplane, to fall within the brightness range of the training data. The transformation increased the brightness of the entire frame using a gamma correction of 1.5 and decreased the brightness within the local object region using a 0.5 gamma correction. An example transformation can be seen in Figure 5.

TABLE II: Comparison between standard robustness techniques

Robustness Technique	Precision	Recall
Baseline	0.970	0.330
Finetuning	0.429	0.238
Traditional Augmentation	0.983	0.307
Targeted Transformation	**0.989**	**0.507**

B. Traditional Robustness Techniques

We compare against two methods from the current literature: fine-tuning and zero-shot transformations. Current zero-shot techniques for improving robustness consist of applying a transformation uniformly to the entire frame [22]. For this comparison, we applied a gamma correction of 1.5 based of our earlier analysis on the image properties of our dataset. For fine-tuning, we started with AirTrack that had been trained on the Airborne Object Tracking [2] dataset and trained it for an additional 500 epochs on the same dataset with the targeted transformation applied.

The results of this comparison are given in Table II. From this, we can see that traditional zero-shot augmentation did not have a significant effect on either the precision or the recall. Fine-tuning, on the other hand, decreased model performance for both metrics. A possible explanation for this is that fine-tuning could have shifted the brightness range that the model was able to successfully detect in, instead of expanding that range. Our method increased the model recall by 53.6%, while keeping the precision high.

C. Targeted Transformations

In this section, we show that the best transformation, based on our analysis of the model dependence on image properties, is a brightness adjustment using a gamma correction of 1.5 on the whole frame and gamma correction of 0.5 in the local object region. This is consistent with the image property measurements we presented in Section IV and is explored in more detail with an ablation study later in this section. The results of this experiment are shown in Figure 6, broken down by each video sequence in the dataset.

The first key aspect of using this technique is to find a useful transformation, as discussed in depth in the previous section. In addition, we perform an ablation study on the construction of this transformation. We test different combinations of gamma parameters, in both the global and local areas, and an orientation transformation, that is mirroring the region about the vertical axis. The results of this ablation study are given in Table III.

These ablations results are consistent with the image property analysis from Section IV. We see that the biggest

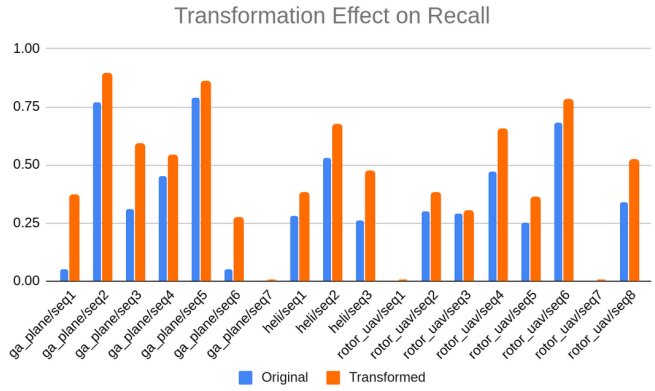

Fig. 6: Model recall results for each video sequence in the TartanX6C dataset with the best transformation: gamma correction of 0.5 in the local object region and 1.5 in the rest of the frame. The recall increases up to 7x depending on the lighting conditions of the particular video.

TABLE III: Ablation results for transformation composition. The baseline performance is given at the top for comparison. The local and global gamma corrections each contribute to a significant increase in recall, giving the best performance when applied together. Flipping the region about the vertical axis gives no improvement in performance, as consistent with earlier analysis.

Transformation Composition				
Global Gamma	Local Gamma	Vertical Flip	**Precision**	**Recall**
1.0	1.0	No	0.970	0.330
0.5	1.0	No	0.887	0.212
0.5	1.5	No	0.857	0.162
1.0	0.5	No	0.977	0.436
1.0	1.5	No	0.963	0.264
1.5	0.5	No	**0.989**	**0.507**
1.5	1.0	No	0.987	0.404
0.5	1.0	Local	0.888	0.213
0.5	1.5	Local	0.858	0.162
1.0	0.5	Local	0.977	0.436
1.0	1.5	Local	0.963	0.264
1.5	0.5	Local	**0.990**	**0.507**
1.5	1.0	Local	0.988	0.404
0.5	1.0	Global	0.342	0.170
0.5	1.5	Global	0.276	0.126
1.0	0.5	Global	0.545	0.372
1.0	1.5	Global	0.367	0.181
1.5	0.5	Global	**0.609**	**0.446**
1.5	1.0	Global	0.543	0.336

improvement is caused by a global gamma correction of 1.5. The next most significant increase in recall is caused by a local gamma correction of 0.5, with these two transformations combined giving the best performance. We also test flipping the transformation region about the vertical axis, which has no effect on performance. This was expected as object orientation had no correlation with performance in our earlier image property analysis.

In addition to constructing the transformations for the global frame and local object region, a critical aspect of this method that will change based on use case is determining the local object region. With this zero shot approach, we see the greatest improvement in performance when the transformed local area regions contain all the objects to be detected.

TABLE IV: Ablation results for local object region selection. The model performance improvement scales with the percentage of objects covered in the transformation regions, with the best performance when all the ground truth (GT) regions are transformed. This method is also robust to incorrect candidate bounding boxes, as the addition of incorrect regions does not affect performance.

Bounding Box Selection	**Precision**	**Recall**
Baseline	0.970	0.330
GT bbox	**0.989**	**0.507**
Learned bbox	0.971	0.329
50% GT + 50% Learned bbox	0.988	0.412
50% GT + Learned bbox	0.987	0.404
GT + 50% Learned bbox	**0.989**	**0.507**
GT + Learned bbox	**0.989**	**0.507**

To test the sensitivity to the bounding box selection, we performed an ablation study on this parameter. We trained an instance of HRNet [40] in the same way as the full frame detector from [12] to learn potential candidate bounding boxes. Then, we tested different combinations bounding box sets, varying the proportion of learned bounding boxes and the proportion of ground truth bounding boxes. We see that, as long as the correct regions are transformed, the method can withstand the addition of many other potentially incorrect regions to this set. Additionally, the performance improvement scales with the percentage of objects covered in the transformation regions. This shows that, when learning these regions to give more attention, it is important to minimize the false negative rate, but not the false positive rate. The results of this ablation study are given in Table IV.

VI. Conclusion & Future Work

We showed that standard techniques employed for improving robustness in generic object detection do not produce the desired results for small object detection, particularly within the application of Detect and Avoid. We presented targeted transformations, a zero-shot technique that can significantly improve robustness with minimal impact on accuracy for detection problems with low signal-to-noise ratio. We demonstrated how to construct these transformations and evaluated our method on the current SOTA model for DAA, showing a 53.6% increase in recall. We also presented an ablation study on the choice of local object regions to augment and the choice of image transformations. In the future, we hope to explore automating the transformation selection process to generalize this method to other domains easily.

Acknowledgments

This work is supported by the Office of Naval Research (Grant N00014-21-1-2110). RM is supported by the National GEM Consortium Fellowship. This work used Bridges-2 at PSC through allocation cis220039p from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #213296, and also supported by a hardware grant from Nvidia.

REFERENCES

[1] Standard specification for detect and avoid system performance requirement. Technical Report ASTM F3442/F3442M-2, ASTM International, 2020.

[2] AICrowd. Airborne object tracking challenge.

[3] Muhammad Waseem Ashraf, Waqas Sultani, and Mubarak Shah. Dogfight: Detecting drones from drones videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7067–7076, 2021.

[4] Ryan Carnie, Rodney Walker, and Peter Corke. Image processing algorithms for UAV sense and avoid. In *Proceedings 2006 IEEE International Conference on Robotics and Automation.*, pages 2848–2853. IEEE, 2006.

[5] Shuaijun Chen, Xu Jia, Jianzhong He, Yongjie Shi, and Jianzhuang Liu. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11018–11027, 2021.

[6] Ying Chen, Xu Ouyang, Kaiyue Zhu, and Gady Agam. Semi-supervised dual-domain adaptation for semantic segmentation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 230–237. IEEE, 2022.

[7] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] Debadeepta Dey, Christopher Geyer, Sanjiv Singh, and Matt Digioia. Passive, long-range detection of aircraft: towards a field deployable sense and avoid system. In *Field and Service Robotics*, pages 113–123. Springer, 2010.

[10] Manuel F Fernandez. Detecting and tracking low-observable targets using IR. In *Signal and Data Processing of Small Targets 1990*, volume 1305, page 193. International Society for Optics and Photonics, 1990.

[11] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven test-time adaptation. *arXiv preprint arXiv:2207.03442*, 2022.

[12] Sourish Ghosh, Jay Patrikar, Brady Moon, Milad Moghassem Hamidi, et al. AirTrack: Onboard deep learning framework for long-range aircraft detection and tracking. *arXiv preprint arXiv:2209.12849*, 2022.

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[14] Cambridge in Colour. Understanding gamma correction. https://www.cambridgeincolour.com/tutorials/gamma-correction.htm.

[15] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022.

[16] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12975–12984, 2020.

[17] John Lai, Jason J Ford, Luis Mejias, and Peter O'Shea. Characterization of sky-region morphological-temporal airborne collision detection. *Journal of Field Robotics*, 30(2):171–193, 2013.

[18] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

[19] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[20] Yang Liu, Peng Sun, Nickolas Wergeles, and Yi Shang. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172:114602, 2021.

[21] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De la Torre. Zero-shot model diagnosis. *arXiv preprint arXiv:2303.15441*, 2023.

[22] Udit Maniyar, Aniket Anand Deshmukh, Urun Dogan, Vineeth N Balasubramanian, et al. Zero shot domain generalization. *arXiv preprint arXiv:2008.07443*, 2020.

[23] Jeffrey W McCandless. Detection of aircraft in video sequences using a predictive optical flow algorithm. *Optical Engineering*, 38(3):523–530, 1999.

[24] Aaron Mcfadyen and Luis Mejias Alvarez. A survey of autonomous vision-based see and avoid for unmanned aircraft systems. *Progress in Aerospace Sciences*, 80:1–17, 2016.

[25] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *Advances in neural information processing systems*, 30, 2017.

[26] Andreas Nussberger, Helmut Grabner, and Luc Van Gool. Aerial object tracking from an airborne platform. In *2014 international conference on unmanned aircraft systems (ICUAS)*, pages 1284–1293. IEEE, 2014.

[27] Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh Ap. Generalization on unseen domains via inference-time label-preserving target projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2021.

[28] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387, 2016.

[29] Anirudha Ramesh, Anurag Ghosh, Christoph Mertz, and Jeff Schneider. Enhancing visual domain adaptation with source preparation. *arXiv preprint arXiv:2306.10142*, 2023.

[30] Vladimir Reilly, Haroon Idrees, and Mubarak Shah. Detection and tracking of large number of targets in wide area surveillance. In *European conference on computer vision*, pages 186–199. Springer, 2010.

[31] Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Flying objects detection from a single moving camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4128–4136, 2015.

[32] Tushar Sangam, Ishan Rajendrakumar Dave, Waqas Sultani, and Mubarak Shah. Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6006–6013. IEEE, 2023.

[33] Falk Schubert and Krystian Mikolajczyk. Robust registration and filtering for moving object detection in aerial videos. In *2014 22nd International Conference on Pattern Recognition*, pages 2808–2813. IEEE, 2014.

[34] FAA Sponsored Sense and Avoid Workshop. Sense and avoid (SAA) for unmanned aircraft system (UAS). In *Final Report of the FAA SAA sponsored workshop*, 2009.

[35] Ankit Singh. CLDA: Contrastive learning for semi-supervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:5089–5101, 2021.

[36] Antonio Tavera, Fabio Cermelli, Carlo Masone, and Barbara Caputo. Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1626–1635, 2022.

[37] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[38] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019.

[39] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European conference on computer vision*, pages 642–659. Springer, 2020.

[40] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation

learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021.

[41] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

[42] Jeongbeen Yoon, Dahyun Kang, and Minsu Cho. Semi-supervised domain adaptation via sample-to-sample self-distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1978–1987, 2022.

[43] Junyi Zhang, Ziliang Chen, Junying Huang, Liang Lin, and Dongyu Zhang. Few-shot structured domain adaptation for virtual-to-real scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[44] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1390–1399, 2021.

[45] Zijian Zhu, Yichi Zhang, Hai Chen, Yinpeng Dong, Shu Zhao, Wenbo Ding, Jiachen Zhong, and Shibao Zheng. Understanding the robustness of 3D object detection with bird's-eye-view representations in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21600–21610, 2023.

[46] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.