Motion Matters: Difference-based Multi-scale Learning for Infrared UAV Detection

Ruian He, Shili Zhou, Ri Cheng, Yuqi Sun, Weimin Tan, Bo Yan*
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University, Shanghai, China
{rahe16, slzhou19, yqsun20, wmtan, byan*}@fudan.edu.cn, rcheng22@m.fudan.edu.cn

Abstract

Unmanned Aerial Vehicle (UAV) detection in the wild is a challenging task due to the presence of background noise and the varying size of the object. To address these obstacles, we propose a novel learning framework for robust UAV detectors, which we call Difference-based Multi-scale Learning (DML). We argue that motion information matters in UAV detection because of the low recognition in one frame. Our method utilizes the frame difference of multiple previous frames, extracting motion information and blocking background noise. We also fuse multiple spatial-temporal scales for training and inferencing, enabling fusion from different sources. In addition, to better evaluate the performance of UAV detection in different scales, we propose Multi-Scale Average Precision (MSAP) metric to aggregate the detection accuracy over multiple scales. Through extensive experiments, we demonstrate that our proposed approach improves the detection accuracy of baseline models. Notably, we achieve SOTA performance in the 3rd Anti-UAV Challenge, with 2nd place in Track 2 and 4th place in Track 1. [1]

1. Introduction

Unmanned Aerial Vehicles (UAVs), commonly known as drones, have become increasingly prevalent in various civil applications [29, 40] due to their flexibility, affordability, and popularity. However, the potential threat [17, 42] they pose to public safety cannot be ignored. UAVs have been used to conduct physical and cyber-attacks and can also violate aviation safety regulations, causing disruptions and economic losses for airlines. In this regard, developing anti-UAV techniques has become a crucial research direction. While radar technology has effectively detected traditional airborne threats, it faces significant challenges in detecting small UAVs due to their low radar cross-sections, erratic flight paths, and low flight altitudes. In contrast, RGB [8, 21] and infrared sensors [1, 36] are well-adopted for small object detection. Compared to RGB sensors, thermal infrared(TIR) sensors are better under extreme conditions, especially for low-light scenes or poor weather.

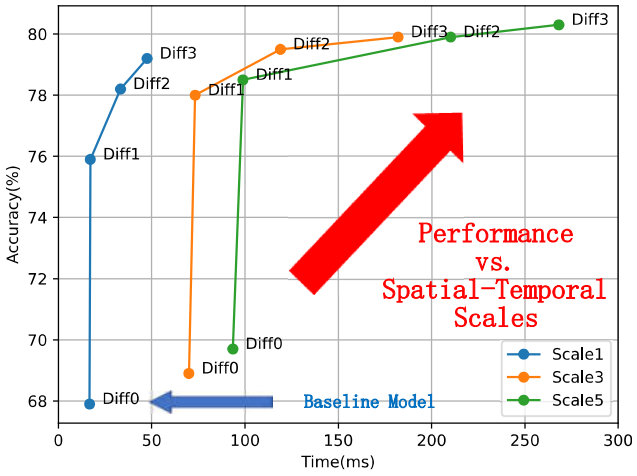
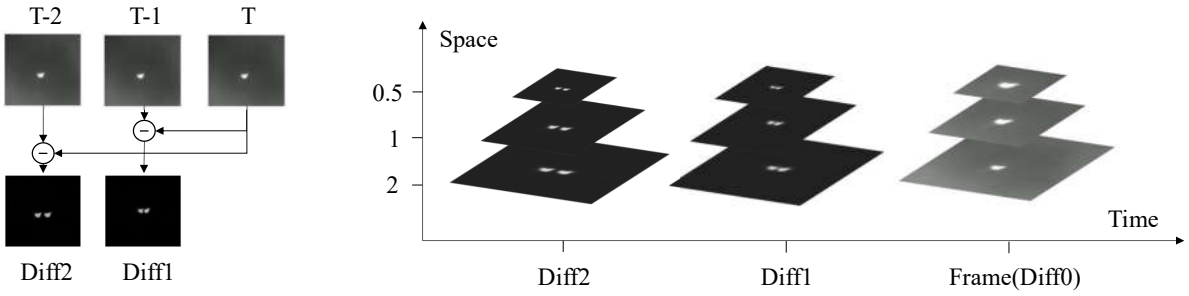


Figure 1. Performance comparison with different spatial-temporal scales. The baseline model is YOLOv5. Diff means the number of frame differences, *i.e.* temporal scales. Scale means the number of spatial scales. The performance improves significantly with spatial-temporal scales, especially when Diff changes from 0 to 1.

Recently computer vision and machine learning algorithms have emerged as promising tools for UAV detection and tracking in TIR [6, 28, 41]. However, two primary challenges remain, background noise and varying target sizes. Background noise can significantly reduce the detection accuracy of UAVs, as it can create false positives and interfere with the detection model. In addition, the varying sizes of UAVs can make them difficult to detect, especially for traditional detection methods that rely on fixed-size priors. Previous works on UAV tracking have highly relied on

[1]This work is supported by NSFC (Grant No.: U2001209, 61902076) and Natural Science Foundation of Shanghai (21ZR1406600).
* Corresponding author: Bo Yan.



(a) Multi-frame Differencing

(b) Spatial-Temporal Scales

Figure 2. A conceptual overview of our proposed method. (a) We use differences from multiple frames as input for the detection model. It makes the detector aware of the motion information. (b) Our model learns with multiple temporal and spatial scales. Fusing different scales will enhance the detector and make it robust to complex scenes.

templates [15, 16, 43], which can be limited in their effectiveness in challenging scenarios with occlusion, noise, and varying object sizes.

We argue that a robust detector should adapt to UAV appearance, motion, and size changes, be robust to occlusions and background noises, and enable the system for more accurate and reliable detection and tracking. To address the problem, we propose a novel learning framework for robust UAV detectors, which we call Difference-based Multi-scale Learning (DML), as shown in Fig. 2.

We propose multi-frame differencing to utilize the motion information in a video sequence. Since the drones are small and have low recognition in noisy backgrounds, more than learning on single frames is needed for robust detection. Frame differencing is commonly used in motion detection [2, 35]. However, the motion information between two frames may be limited, and the moving speed of UAVs is diverse. To address this limitation, we use multiple frames for differencing and enhancing the detection of both slow-moving and fast-moving UAVs. It subtracts the current frame's pixel values from the previous frames' pixel values and highlights the areas where motion has occurred.

We also propose to fuse multiple spatial-temporal scales to improve the detection of objects that vary in size and motion. In the case of UAV detection, the size of the UAV can be significantly different across video frames. Therefore a robust detector needs to be able to handle this variation. Aggregating the detection from different scales makes the detector aware of different speeds and sizes.

To better evaluate UAV detection over diverse sizes, we propose multi-scale average precision (MSAP). Existing single-class mAP or IoU [42] are not appropriate for UAVs because the difficulty is different for different sizes. Therefore, MSAP calculates the detection accuracy over multiple scales. Moreover, we emphasize the accuracy of small objects, which will better reflect real-world performance.

Finally, we develop our method on the 3rd Anti-UAV Challenge Dataset, which enlarges the dataset from previous Anti-UAV competitions [17, 42]. As shown in Fig. 1, extensive experiments demonstrate that our method can significantly improve the accuracy and achieve state-of-the-art performance on the test set. Notably, our model achieved SOTA performance on the dataset and took 2nd place in the Anti-UAV Detection & Tracking track (Track2) and 4th place in the Anti-UAV Tracking track (Track1). Our contributions can be summarized as follows:

1. We develop a novel learning framework for robust UAV detectors, which we call Difference-based Multi-scale Learning (DML). It uses information from multiple spatial-temporal scales to enhance the perception of the detectors.

2. We propose multi-frame differencing to extract motion information from previous frames, which can improve detection accuracy under challenging scenes with occlusion and noise. Moreover, we integrate multiple spatial scales to utilize multiple resolutions, allowing it to detect objects of diverse sizes.

3. We propose a scale-aware metric for UAV detection, which can better evaluate the performance of different sizes of UAVs.

2. Related Work

2.1. Object Detection

Object detection is identifying and locating the presence of objects of interest in an image or video. There are several approaches to object detection, including traditional methods, machine learning, and deep learning. The traditional approach [7, 26] usually has three stages: informative region selection, feature extraction, and classification of the object. Machine learning and deep learning approaches automate these stages by training with annotated data. Deep object detection models can be catego-

rized into two branches: two-staged detectors, represented by RCNN [12] and Mask R-CNN [13], and one-stage detectors, represented by YOLO [31] and SSD [24]. In general, two-stage methods are more accurate than one-stage methods because they explicitly use the RoI Align operation to align an object's features. However, the recent one-stage detectors [10, 37] have narrowed the performance gap and have faster inferencing speed. We implement our approach based on the well-adopted YOLOv5 to show our effectiveness.

2.2. Small Object Detection

Small object detection is a subfield of object detection that focuses on detecting small objects in an image or video feed. Identifying small objects is challenging because small objects often move fast, are occluded by other objects, or have low contrast. It is difficult to locate small objects, especially in a noisy background. There are several techniques used in small object detection. Some of these techniques include increasing image resolution [3, 9], augmenting input data [19, 24], introducing context information [4, 5], and scale-aware training [22, 25, 39]. Our approach fuses the large-range spatial and temporal information for training and inferencing for a given detector, excellently improving the accuracy.

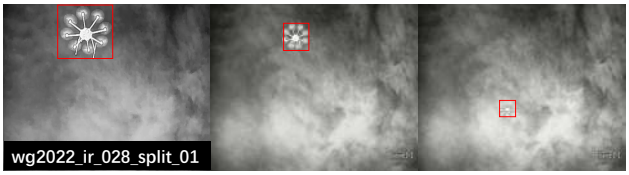
2.3. Infrared Object Detection

Infrared object detection involves detecting objects in an image or video using infrared radiation. Due to the intrinsic of infrared images, the targets usually have a low signal-to-noise ratio and low contrast in a heavily noisy background, which is challenging to detect. Deep learning methods have prevailed in recent works for learning from a large amount of data covering complex scenes. TIRNet [6] adopt a VGG [34] network in a end-to-end manner. McIntosh et al. [28] builds a target-to-clutter network based on Faster-RCNN [11] and Yolo-v3 [32]. ISNet [41] uses a U-Net [33] structure with edge-aware blocks to leverage the edge as a critical feature. However, UAVs differ from other objects in that UAVs are small, and fast-moving, so temporal information is critical in UAV detection and tracking [15, 43]. Our approach uses multi-frame differences to reduce the noise and introduce moving traces.

3. Methodology

3.1. Revisiting Infrared UAV Detection

Infrared images are well-used to detect UAVs due to their ability to detect heat signatures. However, there are several challenges in anti-UAV detection in infrared images. As shown in Fig. 3(a), one of the challenges is the varying size of the UAV. [30] Small UAVs, in particular, are challenging to detect due to their low visibility and limited feature



(a) Varying Size

(b) Disturbing Background

Figure 3. Two main challenges in UAV detection. (a) The drone may fly away from the camera, and its size is drastically changing. (b) Backgrounds such as trees, clouds, and buildings are common distractions that make the drone less recognizable.

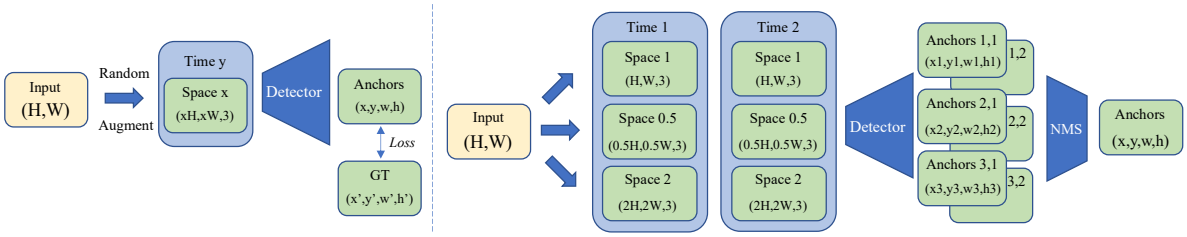
information. At the same time, huge targets are challenging to detect. Because most of the UAV data are focused on smaller scales and some model architectures, like YOLOv5, are based on fixed anchors, the confidence level of huge targets is low.

Another challenge is the low contrast between the UAV and the background, which makes it difficult to detect the UAV in the noisy background. Fig. 3(b) demonstrates the drone in a complex background, such as a building, with very low recognition. It is tough to distinguish the drone's position by just a single image. Additionally, the UAV's shape, orientation, speed, and altitude can also affect the detection performance.

The general detectors like YOLO [31] failed to address the previous problems since the limited ability in spatial and temporal scale awareness. Previous anti-UAV models [20, 25] focus on multiple spatial scale learning for small UAVs. However, the motion also matters for small object learning. Due to the background noise and low visibility for small UAVs, it is hard to detect with only the spatial information. Recent work on UAV tracking [15] also introduces a change detection-based correlation filter which enhances the features with motion information and achieves excellent performance. Therefore, we emphasize the motion information to be as the same important as spatial information.

3.2. Framework Overview

We introduce difference-based multi-scale learning for infrared UAV detection. As shown in Fig. 2, the difference here refers to frame differences, including motion information from previous frames. Multi-scale refers to multiple spatial and temporal scales. Our method is a plug-and-play framework for different detectors. Fig. 4 shows the detailed training and inferencing procedures. We use the popular



Training Strategy

Inferencing Strategy

Figure 4. Training and inferencing strategy of our method. If time is y and space is x, the input is augmented to x ratio resolution with the frame difference between the current frame with the previous y-th frame. And the input is expanded to 3 channels from 1 channel(infrared image) before being fed into the detector.

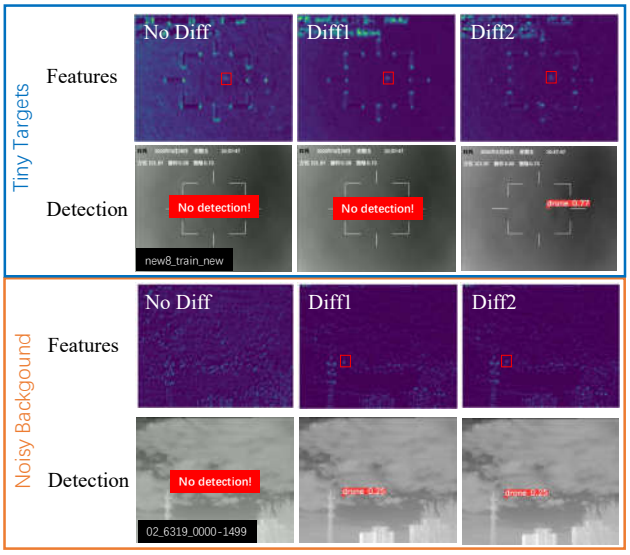


Figure 5. Visualization of multiple temporal scales. We show two examples of tiny targets and noisy backgrounds separately. The features are extracted from the third conclusion layer.

detector YOLOv5 as our base model.

For the training process, we randomly apply spatial and temporal scales for training and expand the infrared input (1 channel) to 3 channels. For the inferencing process, the input is multiplied ($Temporal\ Scales \times Spatial\ Scales$) times for augmentation. Then all scales of inputs are fed into the detector to generate output anchors. Finally, we filter anchors with Non-Maximum Suppression(NMS) and select the most confident one for output.

3.3. Multi-frame Differencing

Temporal information is essential in small object detection, especially for infrared drones. Because of the UAV's small size and fast motion, the IR background has significant background noise and more occlusions in the complex background. The frame difference method is the classical algorithm to remove background noise and extract motion information. Multi-frame frame difference can improve UAV detection because it can adapt to different motion amplitudes of UAVs.

Fig. 2(a) shows the generation of multi-frame differences. Since only previous frames can be utilized in a UAV detection pipeline for streaming videos, our frame differences are calculated with the following formula:

$$Diff_i = I_t - I_{t-i} \tag{1}$$

where $Diff_i$ is the frame difference for previous $i-th$ frame of time t. We feed the frame difference as input to the network with the original image and propose a fusion for frame difference input. The augmented inputs $\mathrm{TAug}_i(I_t)$ can be expressed as follows:

$$\mathrm{TAug}_i(I_t) = [I_t, Diff_i, Canny(Diff_i)] \tag{2}$$

where $Diff$ can be selected from different time intervals i. The input includes the original image, frame difference, and edge as the three channels, and the edge is extracted from the frame difference by the Canny operator. According to [41], infrared images have noisy backgrounds, and the edge information can be more accurately localized for small objects. Therefore, we include canny edges as a part of the input.

Fig. 5 demonstrates the effectiveness of the multi-frame differencing. We visualize the features of the convolution layers with the detection results. With no differencing, the detector suffers from background noise and fails to have a confident prediction. With more temporal information, even slow-moving small objects can also be detected.

3.4. Fusing Multiple Spatial-Temporal Scales

Exploiting spatial information is also a critical technique for small object detection. The tiny object has features of low recognition, and huge objects are also hard to detect for

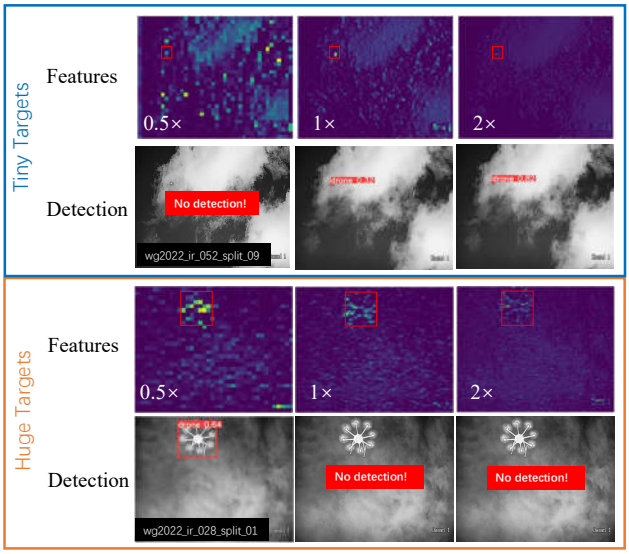


Figure 6. Visualization of multiple spatial scales. We show two examples of tiny targets and large targets separately. The features are extracted from the third convolution layer.

detectors with fixed anchors and trained on a dataset full of small objects. The multi-scale information is widely applied to detection, such as feature extraction in FPN [23] and anchors in YOLO [32]. We here combine multi-spatial scales with multi-temporal scales to enhance the perceptual capability of the network. As shown in the figure, we used joint enhancement during training and testing. During training, our spatiotemporal augmented input can be represented as:

$$\mathrm{STAug}_{i,j} = \mathrm{SAug}_j(\mathrm{TAug}_i(I_t)) \tag{3}$$

where SAug_j is the resolution scaling of j ratio. For example, with a 640×512 input, the SAug_2 means we linearly interpolate the input to 1280×1024.

For inferencing steps, we predict from all kinds of augmented input, do Non-maximum Suppression(NMS) for all anchors, and select the bounding box with the highest score. The prediction process can be expressed as follows:

$$\mathcal{P} = NMS\left(\bigcup_{i,j} \mathcal{F}(\mathrm{STAug}_{i,j}(I_t))\right) \tag{4}$$

where $\mathcal{F}$ is the detection model which will generate the predicted anchors. The final prediction $\mathcal{P}$ is produced by doing NMS over the union of all temporal scale i and spatial scale j.

Fig. 6 demonstrates how spatial scales affect performance. For tiny objects, the low-resolution images such as $0.5\times$ and $1\times$ resolution, the features are noisy, and the detections are not confident. Due to the dataset's bias for small objects, the high-resolution input will have low-recognition features for huge objects. We can achieve robust detection with multiple scales.

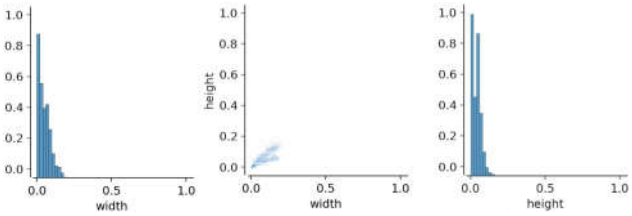


Figure 7. Height and width distribution. UAVs are usually small in height and width, and vary in scale and height-width ratio.

Dataset	Small	Medium	Large	Total
Train	85,087	60,203	18,134	163,424
Validation	29,492	18,518	4,702	52,712

Table 1. Statistics of the 3rd Anti-UAV Challenge Dataset. Small drones have a height or width of under 0.025 ratios of the image; medium drones are under 0.05 and above 0.025; large drones are above 0.05.

3.5. Multi-scale Evaluation

Fig. 7 shows the distribution and correlation of drone sizes in the 3rd Anti-UAV Challenge training set. The size of the drones varies very much, from a few pixels to more than a hundred pixels (0.01 to 0.1 of the original image size) exist, so it is not reasonable to use a uniform evaluation index. Moreover, most drones are of small sizes, which we need to pay more attention to. However, existing single-class metrics are not appropriate for general UAV detection because the size of UAVs is variable. Moreover, the difficulty is different for different sizes of UAVs, resulting in a large gap between the metrics and the real-world detection performance.

We propose the multi-scale average precision (MSAP) metric, which integrates the detection accuracy above multiple scales. Specifically, we divide the UAV into multiple scales based on the proportion of the UAV to the whole image and then calculate the average precision (AP) separately for the results of UAV detection before weighting the average. The input is from a single class of UAVs, and the output is a multi-scale integrated evaluation index. The metric can be expressed as a harmonic mean of the average precision of 3 scales:

$$MSAP = \frac{3}{\frac{1}{AP_s} + \frac{1}{AP_m} + \frac{1}{AP_l}} \tag{5}$$

where AP_s, AP_m, and AP_l indicate the average precision at 0.5 IoU for small, medium, and large UAVs. We set large

Time	Space	Precision	Recall	AP50	AP95	Small	Medium	Large	MSAP	Accuracy	Time(ms)
0x	1x	0.847	0.727	0.801	0.471	0.512	0.928	0.963	0.737	0.679	16.7
0x	3x	0.877	0.735	0.814	0.471	0.546	0.931	0.965	0.761	0.689	69.9
0x	**5x**	**0.883**	**0.742**	**0.817**	**0.477**	**0.549**	**0.934**	**0.968**	**0.764**	**0.697**	**93.5**
1x	1x	0.868	0.75	0.816	0.483	0.546	0.929	0.974	0.762	0.759	17.1
1x	3x	0.897	0.778	0.843	0.485	0.626	0.932	0.971	0.811	0.780	73.3
1x	**5x**	**0.898**	**0.782**	**0.846**	**0.489**	**0.631**	**0.935**	**0.972**	**0.815**	**0.785**	**98.9**
2x	1x	0.87	0.763	0.823	0.485	0.561	0.935	0.974	0.773	0.782	33.3
2x	3x	0.897	0.786	0.846	0.484	0.630	0.936	0.971	0.814	0.795	119
2x	**5x**	**0.898**	**0.791**	**0.85**	**0.489**	**0.637**	**0.940**	**0.972**	**0.819**	**0.799**	**186.8**
3x	1x	0.872	0.766	0.826	0.485	0.566	0.937	0.975	0.777	0.792	47.5
3x	3x	0.896	0.79	0.846	0.483	0.630	0.938	0.971	0.814	0.799	182
3x	**5x**	**0.899**	**0.794**	**0.85**	**0.489**	**0.637**	**0.942**	**0.973**	**0.820**	**0.803**	**268.2**

Table 2. Evaluation on different spatial and temporal scales. The first row indicates the YOLOv5 baseline model. Temporal 0x means no frame difference is used, and Temporal 1 means only one frame difference is used. Temporal 2x and 3x will run the detection for 2 and 3 frame difference inputs for generating final results. Spatial 1x use only the original resolution which is 640×512. And Spatial 2x use [0.5,1,2] resolution ratio for inference, and Spatial 3x use [0.5,0.75,1,1.5,2] with [0.75,1.5] flipped left-right. Small, Medium, Large indicate the AP of the specified size of the drones. Bold texts indicate the best results of the same temporal scale.

objects for sizes > 0.05 of the image, medium objects for sizes 0.05 to 0.025, and small objects for sizes < 0.025. We show the detailed statistics of the training and validation set of the 3rd Anti-UAV Challenge Dataset in Tab. 1. The majority is the small drones which are more challenging to detect. The harmonic mean gives more weight to smaller items and less to larger items to balance the values.

4. Experiments

4.1. Experimental Settings

Dataset Details. The 3rd Anti-UAV dataset is a dataset for discovering, detecting, recognizing, and tracking Unmanned Aerial Vehicle (UAV) targets in the wild and simultaneously estimating the tracking states of the targets given Thermal Infrared (TIR) videos. The dataset is used for the Anti-UAV Challenge [17, 42], a competition for developing algorithms for detecting and tracking UAVs in the wild. The dataset has been released in three subsets, the training subset, the test subset for track 1, and the test subset for track 2. The training subset consists of 200 thermal infrared video sequences and publishes detailed annotation files (whether the target exists, the target location, and many environment labels). The frame size is 640×512, and there are at most 1500 frames in a video. We adopt the official split of the training set for the ablation study, which has 150 videos for train and 50 for validation. Then we train our model on the training and validation subset and test on tracks 1 and 2 for submitting to the 3rd Anti-UAV Challenge.

Evaluation Metrics. Our evaluation metrics are divided into three parts, first is the generic detection metrics, including precision, recall, and average precision(AP); our proposed multi-scale perception metric MSAP; Accuracy (Acc) proposed by 3rd Anti-UAV Challenge. Accuracy is defined as follows:

$$acc = \sum_{t=1}^{T} \frac{IoU_t \times \delta\left(v_t > 0\right) + p_t \times \left(1 - \delta\left(v_t > 0\right)\right)}{T} - 0.2 \times \left(\sum_{t=1}^{T^*} \frac{p_t \times \delta\left(v_t > 0\right)}{T^*} \right)^{0.3} \tag{6}$$

where for frame t, IoU_t is the intersection and union set (IoU) between the predicted tracking frame and its corresponding real frame, and p_t is the predicted visibility flag, which is equal to 1 when the predicted frame is empty and 0 otherwise. v_t is the real label visibility flag of the target. The indicator function $\delta(v_t > 0)$ is equal to 1 when $v_t > 0$ and 0 otherwise. In short, Accuracy metric evaluates the IoU between bounding boxes of the prediction and ground truth, with a penalty on false alerts of drones.

Training Details. We implement our method on the popular YOLOv5 [18] model, a compound-scaled one-stage object detection model. We adopt the YOLOv5 Large model that is pretrained on the COCO dataset as the base model. We train our model for 20 epochs on the training dataset with an SGD optimizer. The batch size is 32, the learning rate is 1e-3. We concatenate the frame difference with the origin frame as RGB channels: the blue channel is the original image, the green channel is the frame difference, and the red channel is the canny edge of the frame difference. We apply multi-scale input for the training batches, for which we randomly resize the input to a resolution ratio from 0.5 to 1.5. The temporal scale y of the training batches is set to

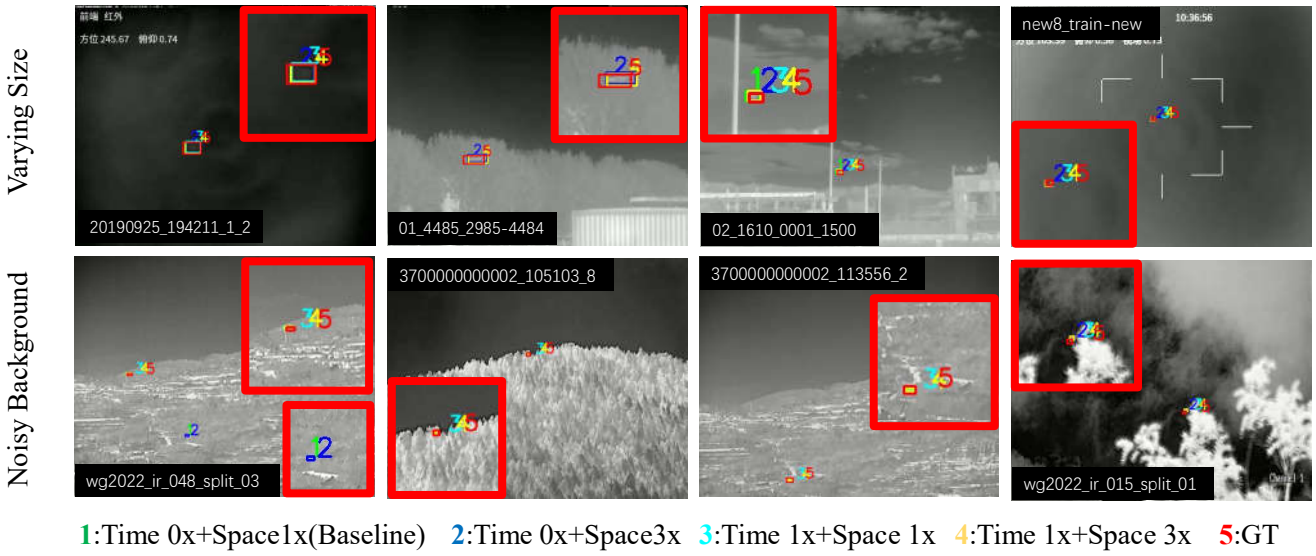


Figure 8. Qualitative comparison on the validation set. We use different colors and digits for different models.

zero when frame differencing is not used (0x in Tab. 2), and one otherwise. We have done ablation on multiple temporal scales in Sec. 4.4, and find that the temporal scale of 1 is better for training.

Inferencing Details. When evaluating on the validation set, we have experimented with different spatial-temporal scale settings (Sec. 4.2), and hyperparameters of confidence and IoU thresholds (Sec. 4.4). The models share the same training settings for spatial scales, which means multiple spatial scale training is applied for each model. And the models differ in the scales of space and time when inferencing. For the models using frame difference, we preprocess the frames with different temporal scales with Eq. (2) and then different spatial scales with Eq. (3). We use the second frame as the previous frame of the first frame because the first frame has no previous frame for frame differencing. We only choose one bounding box with the maximum score for each frame because there is one target in the dataset at most.

4.2. Quantitative Evaluation

Tab. 2 shows the performance and efficiency comparison for each time and spatial scale setting. Our models are trained on the training set for 20 rounds and then tested on the validation set. Time represents the detection time of a frame in milliseconds, and we use one 3090 GPU for the time testing. The first row of which represents the base model of YOLOv5. The model's accuracy grows steadily as the temporal and spatial scales increase. In particular, the model goes from no frame difference to one frame difference, with a 2.2% increase in MSAP, an 8% increase in accuracy, and only a slight increase in time consumption. Finally, we achieved an accuracy of 0.803 and an MSAP of 0.762 on the validation set. Our proposed MSAP is more sensitive to small objects' performance variation than the commonly used Average Precision and can better reflect the real-world detection performance.

At the same time, improving temporal and spatial multiple scales has a powerful impact on the accuracy of small objects. For a model with a spatial scale of 1, boosting the temporal scale can improve the AP by 3.4%, 1.5%, and 0.5%, respectively, while for a model with a temporal scale of 1, boosting the spatial scale can improve it by 8% and 0.5%, respectively. We also found that for a temporal scale of 0 (*i.e.*, no frame difference is used), the boost in spatial scale is less than that for a temporal scale of 1. Therefore, the temporal and spatial scales are correlated, and boosting the temporal scale can also make spatial perception more effective.

4.3. Qualitative Evaluation

Fig. 8 shows the detection results of different methods in two complex cases, including small objects and background noise. Fusing the temporal and spatial scales can detect more small objects. Moreover, increasing the temporal scale is very effective for scenes with complex backgrounds. In particular, for the scene, *wg2022_ir_048_split_03*, the model without frame difference tends to falsely detect other locations as drones, while the model with frame difference will detect the moving drones.

Experiment	Method	Accuracy
Training		
Temporal Scale	Frame+Frame+Frame	0.679
	Frame+Diff1+Canny	**0.759**
	Frame+Diff1+Diff2	0.748
Spatial Scale	No	0.690
	Yes	**0.759**
Inferencing		
IoU Thres.	**0.1**	**0.759**
	0.2	0.756
	0.3	0.755
	0.4	0.755
Conf. Thres.	0.1	0.758
	0.2	**0.759**
	0.3	0.750
	0.4	0.734

Table 3. Ablations on alternative settings. Bold texts indicate the best result.

(1) Wrong Detection

(2) False Alert

Figure 9. Failure cases of our model. Complex background will affect the detector with noise and similar objects.

4.4. Ablation Study

Tab. 3 shows the ablation experiments with alternative hyperparameters and settings other than the multiple spatial-temporal scales. First is the temporal scales used for training. Our Frame+Diff1+Canny input is better than both single frame input and three frame difference input because the edges of small objects can provide more easily identifiable information. Secondly, the multi-scale training can improve the robustness for detecting different size targets and achieve higher accuracy. Finally, the thresholds in NMS and the score (confidence) thresholds for detection also have an impact on the detection performance, and adjusting these thresholds is also a crucial step to improve the accuracy.

5. Discussion

5.1. Limitations

Fig. 9 shows our failure cases. The hard cases usually have a complex background with noise and similar objects, like windows, towers and birds. The detector will be misled if the counterpart is more significant in the view than target UAVs. This can be partly solved with a tracking algorithm to find the most time-consistent target rather than the most significant object.

Fig. 1 and Tab. 2 demonstrate that the time consumption grows linearly with time and spatial scale. Running the model with $3\times$ the temporal scale and $5\times$ times the spatial scale takes 268 ms per frame. To address this drawback, we can increase the parallelism to improve the GPU usage or optimize the execution by introducing techniques such as model compilation to increase the speed. When deploying, we can replace the backbone with a more lightweight model, such as YOLOv5 small or MobileNet [14].

5.2. Future Work

Recent works [27, 38] propose unified models to solve detection and tracking tasks, breaking the routine of detection-by-tracking or tracking-by-detection. Detection can achieve better performance with template prior, motion information, and background estimation by tracking, while tracking relies on the results of detection. How to integrate our robust detector into a unified detection-tracking system is a question worth exploring. Exploiting the history information by combining the detector with a tracking algorithm may give better results for detection.

6. Conclusion

In this paper, we propose a novel learning framework for robust UAV detectors, which we call Difference-based Multi-scale Learning (DML). Our method utilizes the frame difference of multiple previous frames, extracting motion information and blocking background noise. We also fuse multiple spatial-temporal scales for training and inferencing. In addition, we propose Multi-Scale Average Precision (MSAP) metric to evaluate the performance of UAV detection in different scales better. Through extensive experiments, we demonstrate that our proposed approach improves the detection accuracy of baseline models.

References

[1] Amanda Berg, Jörgen Ahlberg, and Michael Felsberg. Channel coded distribution field tracking for thermal infrared imagery. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1248–1256, 2016. 1

[2] Christoph Briese, Andreas Seel, and Franz Andert. Vision-based detection of non-cooperative uavs using frame differencing and temporal filter. *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 606–613, 2018. 2

[3] Zhaowei Cai, Quanfu Fan, Rogério Schmidt Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. *ArXiv*, abs/1607.07155, 2016. 3

[4] Chenyi Chen, Ming-Yu Liu, Oncel Tuzel, and Jianxiong Xiao. R-cnn for small object detection. In *Asian Conference on Computer Vision*, 2016. 3

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2016. 3

[6] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51:1244 – 1261, 2020. 1, 3

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:886–893 vol. 1, 2005. 2

[8] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4655–4664, 2018. 1

[9] Cheng-Yang Fu, W. Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. Dssd : Deconvolutional single shot detector. *ArXiv*, abs/1701.06659, 2017. 3

[10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[12] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2013. 3

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[14] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017. 8

[15] Bo Huang, Junjie Chen, Tingfa Xu, Ying Wang, Shenwang Jiang, Yuncheng Wang, Lei Wang, and Jianan Li. Siamsta: Spatio-temporal attention based siamese tracker for tracking uavs. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1204–1212, 2021. 2, 3

[16] Kutalmis Gokalp Ince, Aybora Koksal, Arda Fazla, and Aydin Alatan. Semi-automatic annotation for visual object tracking. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1233–1239, 2021. 2

[17] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Qixiang Ye, Jianbin Jiao, Zhenjun Han, et al. Anti-uav: a large-scale benchmark for vision-based uav tracking. *T-MM*, 2021. 1, 2, 6

[18] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, Zeng Yifu, Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, Nov. 2022. 6

[19] Mate Kisantal, Zbigniew Wojna, Jakub Murawski, Jacek Naruniec, and Kyunghyun Cho. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019. 3

[20] Aybora Koksal, Kutalmis Gokalp Ince, and Aydin Alatan. Effect of annotation errors on drone detection with yolov3. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4439–4447, 2020. 3

[21] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4277–4286, 2018. 1

[22] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6053–6062, 2019. 3

[23] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2016. 5

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 3

[25] Ziming Liu, Guangyu Gao, Lin Sun, and Lingyu Fang. Ipg-net: Image pyramid guidance network for small object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4422–4430, 2019. 3

[26] G LoweDavid. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 2

[27] Fan Ma, Mike Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, and Zhicheng Yan. Unified transformer tracker for object tracking. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8771–8780, 2022. 8

[28] Bruce McIntosh, Shashanka Venkataramanan, and Abhijit Mahalanobis. Infrared target detection in cluttered environments by maximization of a target to clutter ratio (tcr) metric using a convolutional neural network. *IEEE Transactions on Aerospace and Electronic Systems*, 57:485–496, 2021. 1, 3

[29] Matthias Mueller, Neil G. Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European Conference on Computer Vision*, 2016. 1

[30] Kun Qian, Shenghui Rong, and Kuanhong Cheng. Anti-interference small target tracking from infrared dual waveband imagery. *Infrared Physics & Technology*, 2021. 3

[31] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2015. 3

[32] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018. 3, 5

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 3

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[35] Nishu Singla. Motion detection based on frame difference method. *International Journal of Information & Computation Technology*, 4(15):1559–1565, 2014. 2

[36] Vijay Venkataraman, Guoliang Fan, Joseph P. Havlicek, Xin Fan, Yan Zhai, and Mark B. Yeary. Adaptive kalman filtering for histogram-based appearance learning in infrared imagery. *IEEE Transactions on Image Processing*, 21:4622–4635, 2012. 1

[37] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 3

[38] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. Omnitracker: Unifying object tracking by tracking-with-detection. *ArXiv*, abs/2303.12079, 2023. 8

[39] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13658–13667, 2021. 3

[40] Hongyang Yu, Guorong Li, Weigang Zhang, Qingming Huang, Dawei Du, Qi Tian, and N. Sebe. The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *International Journal of Computer Vision*, 128:1141–1159, 2019. 1

[41] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haicheng Bai, Jing Zhang, and Jie-Ru Guo. Isnet: Shape matters for infrared small target detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 867–876, 2022. 1, 3, 4

[42] Jian Zhao, Gang Wang, Jianan Li, Lei Jin, Nana Fan, Min Wang, Xiaojuan Wang, Ting Yong, Yafeng Deng, Yandong Guo, et al. The 2nd anti-uav workshop & challenge: methods and results. *arXiv preprint arXiv:2108.09909*, 2021. 1, 2, 6

[43] Jinjian Zhao, Xiaohan Zhang, and Pengyu Zhang. A unified approach for tracking uavs in infrared. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1213–1222, 2021. 2, 3