remote sensing

MDPI

Article

HCA-YOLO: A Hierarchical Cross-Scale Attention Learning Framework for UAV Detection

Wei Wang [1,†], Yan Zhang [2,†], Yaxiu Zhang [2], Lingjun Zhao [2] and Xingwei Yan [2,*]

[1] College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, China
[2] College of Electronic Science, National University of Defense Technology, Changsha 410073, China
* Correspondence: yanxingwei@nudt.edu.cn
† These authors contributed equally to this work.

Highlights

What are the main findings?

- A hierarchical cross-scale attention learning framework was proposed for UAV detection in vision image, by which the task of drone detection in adverse conditions can be achieved.
- The hierarchical attention backbone, the quasi-residual neck, and the dynamic focusing measurement were conjoined to promote learning effectiveness.

What are the implications of the main findings?

- The proposed method is capable of detecting drones with mutable scales by the hierarchical attention learning mechanism.
- The drone detection performance in adverse weather conditions is improved significantly by the proposed method.

Abstract

The accurate detection of unmanned aerial vehicles (UAVs) in various sizes played an important role in the practical applications. Yet the preceding works suffered from the missing inference, the false alarms, and the poor accuracy due to the the adverse scene conditions, as well as the mutable scales. To solve the problems, a hierarchical attention promoted cross-scale learning framework was proposed in this paper. First, the hierarchical attention mechanism was introduced in the backbone to generate the multi-scale features of targets, so they can be discerned and located at different scales. The resulting features were further delivered to the neck, in which two branches of features were built, respectively. The former was obtained by the target-specific feature operator, while the latter was generated by the upsampling operation. The dual branches were further connected in the quasi-residual structure. So the content of targets can be protected well, and the detail information can be reconstructed. Finally, the dynamic focusing loss measurement was presented to regress the bounding box of the target, so the learning effectiveness of presented the architecture can be promoted. To verify the proposed method, multiple rounds of experiments were performed. The results demonstrated that small and weak drones can be detected accurately, especially in adverse lighting and weather conditions. The evaluation metric of mean average precision rate (mAP) can be improved by 18.5% (YOLO6) on the collected dataset.

Keywords: anti-UAV; target detection; attention mechanism; cross-scale learning

Check for updates

Academic Editor: Taejung Kim

Received: 14 April 2026
Revised: 10 June 2026
Accepted: 15 June 2026
Published: 5 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the development of microelectronic technology, more and more unmanned aerial vehicles (UAVs) emerged in many fields. The low cost of manufacturing for various kinds of drones made them widely used in many applications. Yet the rapidly increased number of UAVs presented many problems. The drones' low flight altitude, high maneuverability, and stealthiness make them susceptible to misuse, and hence they pose great threats to public and national security. Therefore, it is important to accurately detect the unauthorized "rogue drones" in low-altitude airspace. It is the primary step to take corresponding countermeasures.

1.1. Model-Driven Method

The early studies for UAV detection in vision images relied on the conjunction of handcrafted features and simple learning techniques. It can be generally separated into three sequential steps: region proposal, feature extraction, and location regression. The representative features include scale-invariant feature transform (SIFT) [1], histogram of gradient (HoG) [2], wavelet transform [3]. Many kinds of improved versions were then presented in the preceding works, such as speed-up robust features [4], Root-SIFT [5], DSP-SIFT [6], L2-SIFT [7], and AB-SIFT [8]. The low-level features were then fed into a simple learnable classifier for classification and regression, such as support vector machine, random forest, and logistic regression. This family of method is pictorially shown in Figure 1. The most powerful performance was obtained by the deformable part-based model [9].

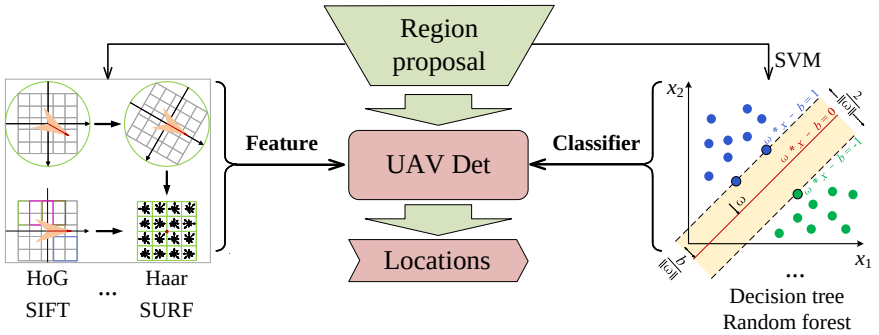


Figure 1. The illustration of classical method for UAV detection.

Though widely studied in the preceding works, the model-driven method still suffered from some typical issues.

- The implementation of these methods was much more complex due to the generation of handcrafted feature during target detection. It required many more procedures, such as keypoint detection, multi-scale space, feature representation. Yet each of these steps was important for the detection performance.
- The detection performance was not robust in practice. It was dependent on the discretization error of the grid, the initial assumption of fuzzy errors, the image filtering.
- The generalization ability to specific image scenes was much poorer, such as scenes with extreme changes in brightness, high-frequency texture areas, and defocused blurry areas. In these scenes, it was very difficult to extract enough features for target detection.

1.2. Learning-Based Strategy

With the development of the theories of deep learning, recent studies for UAV detection gradually resort to learning the high-level features from the input data adaptively. Wu Xuan et al. designed a concatenation and fusion structure that upsamples the shallow feature maps for YOLO4 (You Only Look Once) [10]. The semantic information can be

exploited fully, and the detection performance for UAVs at small scale can be improved. Hao Jinyuan et al. presented the OD-YOLO to address the issues of poor robustness, low accuracy, and high model complexity in drone detection [11]. They reconstructed the downsampling structure and backbone network by incorporating multiple convolutional variants. Wei Feng et al. proposed a lightweight object detection model suitable for mobile devices based on YOLOv5 [12]. The original backbone was replaced by ShuffleNetv2 [13] to reduce model parameters. The SE-attention [14] mechanism was considered in NMS to compensate the accuracy loss. Q. Cheng et al. simplified the feature extraction process using the CF2-MC backbone network and proposed the MG feature fusion module to ensure strong feature representation capability and avoid excessive information loss [15]. Y. Yuan et al. presented a distributed edge-cloud collaborative framework featuring an edge-embedded lightweight algorithm with an attention mechanism [16]. By employing a fuzzy neural network for adaptive task allocation between edge and cloud, the framework optimizes the real-time detection of ground moving targets. This approach achieves superior speed and accuracy compared to traditional single-platform computing on embedded devices. Y. Tao et al. proposed a real-time object detection network that utilizes a feature fusion module and a lightweight backbone [17]. By integrating a convolutional multihead self-attention mechanism and an attention prediction head, the model effectively captures context to identify small and occluded objects. This design ensures high detection accuracy while maintaining efficient real-time performance on embedded platforms. Z. Jiao et al. proposed an object detection model based on the Transformer architecture [18]. By introducing a 2-D Gaussian probability density function and a denoising training method, the model accelerates target localization and stabilizes training. This effectively improves the detection performance for small targets in complex marine environments. Z. Zou et al. developed an active object detection method that integrates shallow FPN features with detector output bounding boxes [19]. By utilizing a two-stage learning strategy involving behavior cloning and Q-learning, the model dynamically adjusts viewing angles and scales. This effectively addresses multiscale targets and occlusions, improving both detection accuracy and inference efficiency in UAV remote sensing.

The development of learning-driven detection methods can be summarized as two kinds of typical architecture: convolutional network and transformer. They are reviewed in Figure 2.

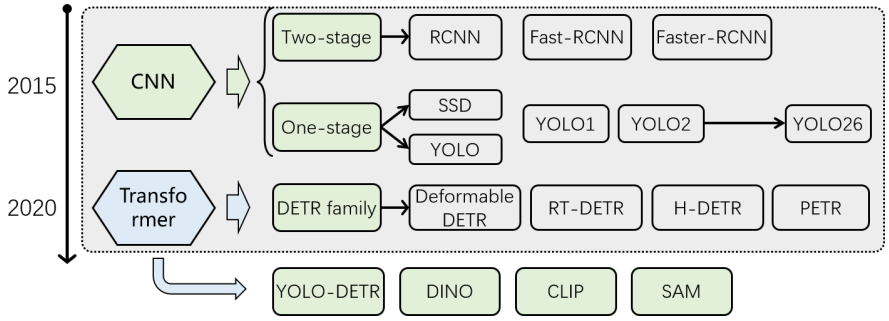


Figure 2. The development of learning-driven detection strategies.

Although the learning-driven methods achieved good performances in certain fields, they still face some challenges.

- The mutable scales. In practical applications, UAVs ranged from several pixels to a hundred pixels in size. The small ones were difficult to discern because the powerful features were unavailable.

- The distant targets. A distant drone usually occupies a few pixels in an image. The details are blurred, leading to a significant loss of physical structure. It is therefore difficult to extract the distinctive local features, such as texture, shape, or key components.
- The adverse environments. Extreme weather and environmental conditions also create significant interference for drone detection, such as poor image quality, low lighting conditions, and interference with similar characteristics. In these conditions, false alarms and missing detection rates were increased [20–24].

1.3. Our Solution

To address the problems above, this paper proposed a new hierarchical attention promoted cross-scale learning framework. Different from the preceding works, both the backbone and the neck were adapted to the task of UAV detection. For the backbone, the hierarchical cross-channel attention mechanism was introduced to generate the multi-scale features of targets so they can be discerned and located at different scales. For the neck, two branches of features were configured, respectively. The former was obtained by the target-specific feature operator, while the latter was generated by the upsampling operation. The dual branches were then connected in the quasi-residual structure. The multi-scale features resulting from the backbone can be aggregated and combined. So the content of targets can be protected well, and the detail information can be reconstructed. Finally, the dynamic focusing loss measurement was presented to regress the bounding box of the target. So the incorrect gradients for drone detection can be eliminated, and hence promote the learning effectiveness of the presented architecture accordingly. The overall architecture of the proposed method was illustrated in Figure 3. To verify the proposed method, multiple rounds of experiments were performed on the collected dataset. The results demonstrated that the proposed method could improve the detection performance of small UAVs, especially in adverse lighting and weather scenes. The evaluation metric of mAP can be improved by 18.5% in comparison to the baseline algorithm (YOLO6).

The major contributions of this paper can be summarized twofold.

- The hierarchical attention learning mechanism deployed network. A hierarchical cross-channel attention mechanism is presented to generate the multi-scale features of targets.
- The dynamic focusing loss measurement strategy. A dynamic focusing loss measurement is presented to regress the bounding box of the target so the incorrect gradients for UAV detection can be eliminated.

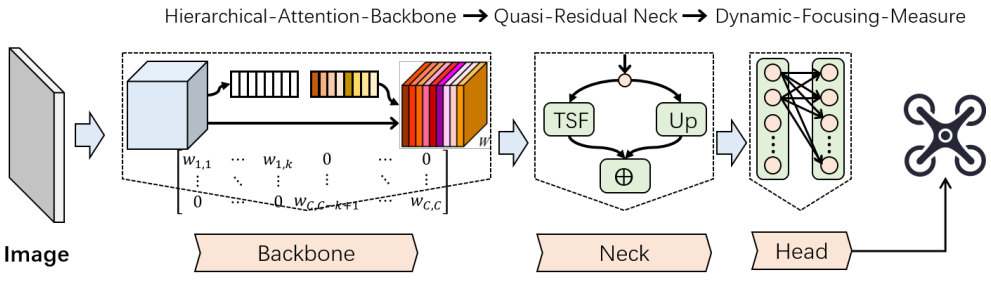


Figure 3. The overall architecture of proposed method.

The reminder of this paper is organized as follows. The general baseline of object detection is described in Section 2. The proposed method is delineated in Section 3. The attention-induced backbone, the quasi-residual neck, and the dynamic focusing loss measurement are presented sequentially. The experiments and the analysis are shown in Section 4, while Section 5 concludes this paper and outlines directions for future research.

2. The Baseline for Target Detection

UAV detection aims to identify and locate various kinds of UAV from the input image. The most popular method for the object detection task is the YOLO detector. It can achieve a trade-off between low computational cost and detection performance. YOLO is a single-shot (one-stage) object detection architecture that performs object localization and classification in a single forward pass through a single neural network. The unified architecture enables the global inference across the input image, leading to fewer false positives and more robust detections, even across different domains. The family of detector has multiple versions, each of which aims to improve the accuracy while maintaining the speed. YOLO follows a simple but powerful pipeline for object detection. The overall architecture includes backbone, neck, and head. It is illustrated in Figure 4.

- Preprocessing. The input image is resized to 448 $\times$ 448 pixels for consistency.
- Grid division. The image is divided into 7 $\times$ 7 grid, each of which is responsible for detecting objects whose center falls inside that cell.
- Feature extraction. A convolutional neural network with 24 convolutional layers followed by 2 fully connected layers is used to extract features from the image.
- Prediction. Each grid cell predicts 2 bounding boxes, including the parameters of (x, y), width, height, confidence. In each grid cell, the class probabilities for each object category are predicted. The final output is a 7 $\times$ 7 $\times$ 30 tensor (7 $\times$ 7 grid $\times$ [2 boxes $\times$ 5 values + classes]).
- Post-processing. The class scores are computed by multiplying class probabilities with box confidence. Non-Maximum Suppression (NMS) is used to eliminate overlapping boxes and produce the final detections.

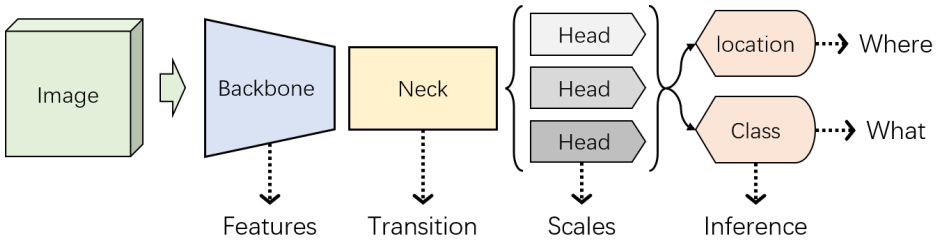


Figure 4. The standard pipeline of YOLO detection algorithm.

Though the YOLO algorithm achieved great performance in certain fields previously, it still faces some challenges for the task of UAV detection due to the following issues.

- The mutable scales of target. In the vision image, the size of UAV targets can range from several pixels to hundreds of pixels, depending on the range of the UAV. So it is difficult to discern and locate with the generic model. In addition, changes of scale between nearby and distant UAVs pose severe challenges to the model's multi-scale adaptation capability.
- The extremely small size of the target. In some conditions, UAVs typically occupy only dozens of pixels in images. It is therefore very difficult to extract the effective features from the input image for target detection.
- The low inter-class difference. The visual features of UAVs are very similar to moving targets like birds and falling leaves in gloomy conditions, such as dark and haze. In addition, it is much more difficult to identify UAVs accurately due to motion blur and occlusion. These difficulties easily produce false alarms and missed detections.

Some exemplars are displayed in Figure 5.

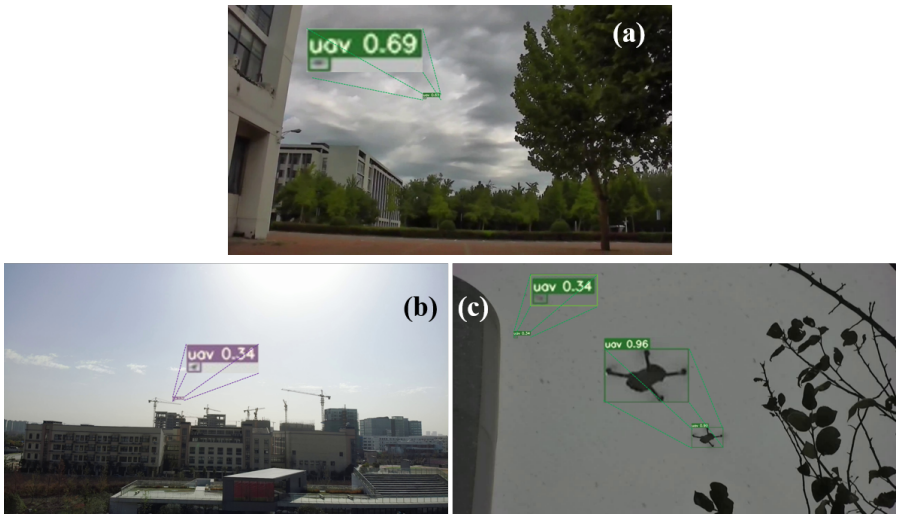


Figure 5. The typical challenges for UAV detection: (**a**) the target's small size, (**b**) the scale variations, and (**c**) the interference.

3. The Proposed Method

To handle UAV detection, a hierarchical attention promoted cross-scale learning framework is proposed in this paper. Different from the generic strategy, the hierarchical attention mechanism is first introduced to build the multi-scale representations of targets. They are used to form the residual connected neck. On this basis, the dynamic focused loss measurement is presented to promote the learning effectiveness. The detail hierarchy of the proposed method is shown in Figure 6.

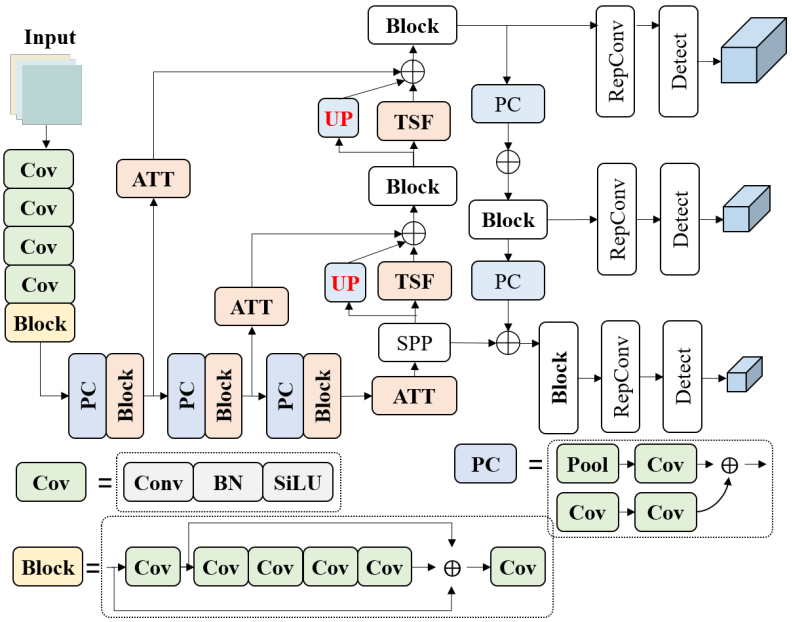


Figure 6. The detail hierarchy of proposed method.

3.1. Hierarchical-Attention Backbone

In the baseline of object detection (see Figure 4), the high-level features are obtained from the backbone, where multiple combinations of convolution blocks are configured. Yet the process flow is not effective for the task of UAV detection. In this paper, the hierarchical attention learning mechanisms are presented to enhance the feature learning further. Three groups of attention hierarchies are deployed on the backbone to generate the representations at different scales, so much more information of the UAV target can be exploited.

3.1.1. Local Attention Mechanism

It was concluded that dimensionality reduction was ineffective and unnecessary to capture the cross-channel dependencies [25–27]. To solve the problem, an efficient channel attention hierarchy is presented. It could preserve the spatial dimensionality of feature tensors. Given the initial feature $\mathbf{y}$, the channel attention is learned by

$$\boldsymbol{\omega} = \sigma(\mathbf{W}\mathbf{y}) \tag{1}$$

where $\mathbf{W}$ is the parameters matrix with $C \times C$ in size, and σ is the sigmoid activation function. To capture the cross-channel interaction, a band matrix $\mathbf{W}_u$ is used:

$$\begin{bmatrix} w_{1,1} & \cdots & w_{1,k} & 0 & \cdots & \cdots \\ 0 & w_{2,2} & \cdots & w_{2,k+1} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{2}$$

where $w_{i,j}$ is the entry of the band matrix. It involves the number of learned parameters $k \times C$. So the weight of y_i is only related to the interaction with the k neighbors:

$$\omega_i = \sigma\left(\sum_j w_i^j y_i^j\right), \quad y_i^j \in \mathcal{N}_i \tag{3}$$

where $\mathcal{N}_i$ is the neighbor channels of y_i, and w_i^j is the weight coefficient. So those UAVs with much smaller scale can be represented efficiently by capturing the local cross-channel interaction.

3.1.2. Kernel Determination

With the operation of global average pooling, the 1-D aggregated features can be formed. The channel weights are then generated by applying a fast one-dimensional convolution with kernel size k. These weights were further mapped to the range of $(0, 1)$ through the sigmoid activation function. The local cross-channel interaction can therefore be achieved. Finally, the weights were multiplied with the original input feature map to produce the enhanced output. Here, the value of kernel size k was adaptively determined. It was positively correlated with the channel dimension C:

$$C = \phi(k) = 2^{(\lambda k - b)} \tag{4}$$

where λ and b are the constants determined experimentally to be 2 and 1, respectively. The presented attention mechanism is illustrated in Figure 7.

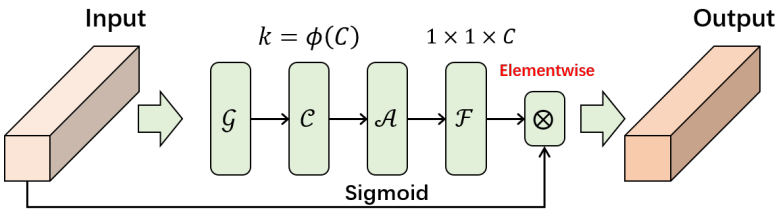


Figure 7. The illustration of presented attention mechanism, where $\mathcal{G}$ denotes the global average pooling, $\mathcal{C}$ is the convolution, $\mathcal{A}$ is the sigmoid activation function, $\mathcal{F}$ is the fully connected layer, and $\otimes$ is the element-wise product.

Given the channel dimension C, the value of k can be derived as follows:

$$k = \psi(C) = \left| \frac{\log_2(C)}{\lambda} + \frac{b}{\lambda} \right|_{odd} \tag{5}$$

where $|\cdot|_{odd}$ is the operation to get an odd number.

Compared with the original attentions, the presented technique is light-weighted and flexible. It could select the size of the convolution kernel k adaptively by means of the local cross-channel strategy, so the model complexity can be reduced significantly. Likewise, much more effective features can be extracted for the subsequent task of UAV detection.

3.2. Quasi-Residual Neck

To handle the mutable size of UAVs, the multi-scale backbone induced by the hierarchical attention mechanisms were presented previously, so the representations of UAV targets at different scales can be obtained. On this basis, the quasi-residual connection technique is developed to promote the representation power in the neck module. One of the branches is formed by the target-specific feature generator (TSF), while the other is the feature upsampling operator (FUO). They are concatenated in the quasi-residual connection, so the learning efficiency can be promoted.

3.2.1. Overall Work-Flow

Since the UAVs in the practical scenarios can be at various kinds of scales, ranging from several pixels to dozens, it is necessary to learn the features at different scales. The traditional upsampling methods include linear interpolation and deconvolution. The former employed simple mathematical interpolation, while the latter utilized reverse convolutions to enhance the feature map resolution. These methods often produced blurred features and artifacts [28]. Detail information was then lost, hence impacting UAV detection.

To solve the problem, a target-specific features generator is presented in this paper. Different from the preceding works, it pays more attention to the actual content of the target feature during upsampling, rather than simple spatial interpolation. In specific, TSF analyzes the low-resolution feature maps to extract the crucial spatial information, and then achieves the adaptive upsampling. The transformation efficiency from the low-resolution feature maps to the high-resolution images can therefore be improved. A compact convolutional layer was used to generate target-specific kernels, which captured the important information from the input features. The high-resolution features can be reconstructed from the input low-resolution ones, during which the contextual information was considered. So the details and edge information can be preserved, while the blurring and artifacts emerging in the traditional methods were circumvented accordingly. So TSF is composed of two components, the upsampling kernel prediction and the feature reorganization. A unique kernel for each target location was predicted according to its content. They were further used to reassemble the features.

- Upsampling kernel prediction. Given the feature of target of size $H \times W \times C$, the specific reorganization kernels were first predicted:

$$W_{n'} = \Psi(\mathcal{N}(X_n, k_{cov})) \tag{6}$$

  where $\mathcal{N}(X_n, k)$ denotes the $k \times k$ neighborhood centered at position n in X, Ψ is the upsampling kernel prediction mapping. The kernel size of the convolutional layer in the encoder was set as k_{cov}, and the reorganization kernel size was set as k_{up}. So a reorganization kernel W_n for each location n' based on the neighborhood of X_n can be estimated.
- Feature reorganization. The predicted kernels were then used for feature reorganization. So the features of target of size $\sigma H \times \sigma W \times C$ can be produced:

$$X'_{n'} = \phi(\mathcal{N}(X_n, k_{up}), W_{n'}) \tag{7}$$

where ϕ is the feature reorganization mapping. Likewise, the high-resolution features can be obtained by the dot product between the neighborhood of X_n and the corresponding reorganization kernel W_n.

3.2.2. Target Specific Features

The target-specific high-resolution features were generated from the low-resolution ones, so the spatial relationships and semantic information among features can be exploited better. The generated features provided a solid foundation for the subsequent feature reorganization, and ensured accurate pixel-to-position mapping during upsampling. To delineate the presented TSF, a box diagram is shown in Figure 8. It consists of three sub-modules: channel compressor, content encoder, and kernel normalizer. The channel compressor performed convolution without altering the size of features. So the number of feature channels can be compressed. A convolution with kernel size $k_{encoder}$ was then applied for content encoding. This produced reorganization kernels with an output channel size of $\sigma^2 k_{up}^2$. The upsampling kernels were unfolded along the spatial dimension. The rearranged kernels had a size of $\sigma H \times \sigma W \times k_{up}^2$. A softmax function was applied to normalize the reorganized upsampling kernels, yielding the adaptive reorganization kernels. Compared with traditional methods, the presented technique could circumvent the blurring and the detail loss caused by simple interpolation or deconvolution operations.

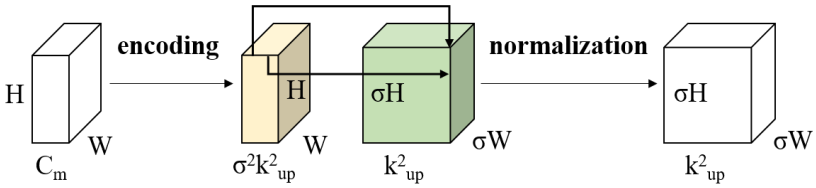


Figure 8. The illustration of kernel prediction during upsampling.

3.2.3. Feature Upsampling Operator

The feature reorganization served as the core component of TSF. It was responsible for transforming low-resolution feature maps into high-resolution images through effective information reorganization and reconstruction. The work flow is shown in Figure 9.

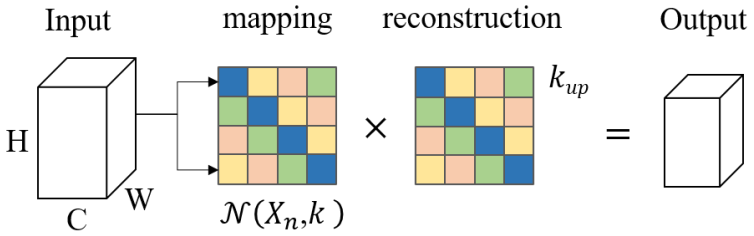


Figure 9. The illustration of feature reorganization.

The convolutional operations were first used to generate a target-specific map. It helped to capture the spatial information and semantic relationships across different regions of the target. Given this feature map, the low-resolution features can then be dynamically reorganized into the high-resolution form. By adapting to the specific content of local regions, the feature reorganization process was then optimized. The global information from the low-resolution features was considered, ensuring that the upsampled results contained more precise details, sharper edges, and richer textures.

3.3. Dynamical-Focusing Measurement

The loss measurement played an important role for the performance of object detection. In the previous works, many kinds of loss functions were presented. The complete IOU (CIOU) loss function was proposed to calculate the bounding box regression loss [29]. On this basis, the distance IoU loss was presented to incorporate the additional scale loss for detection boxes by introducing width and height loss [30]. The predicted boxes can then fit the ground truth boxes better. However, these measurements may not be suitable for UAV detection, because the low-quality geometric factors in the data could increase the penalties from low-quality examples during training [31]. The model generalization ability can therefore be reduced. Consequently, they demonstrated relatively poor detection capability for small targets. To solve the problem, a dynamical focusing loss measurement was presented in this paper. Different from the previous works, the anchor box quality was evaluated using an outlier measurement and constructed distance attention. So the penalties for high-quality anchor boxes can be reduced.

3.3.1. Base Measurement

The CIoU loss function was expressed as

$$\ell_{CIoU} = \ell_{IoU} - \left(\frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \right) \tag{8}$$

where α and v are the aspect ratio parameters, w, h are the width and height of the predicted box, w^{gt}, h^{gt} are the width and height of the ground truth box, and IoU indicates the intersection over union between the predicted and ground truth boxes. The aspect ratio parameters were defined as

$$\begin{cases} v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \\ \alpha = \frac{v}{(1 - Iou) + v} \end{cases} \tag{9}$$

3.3.2. Presented Measurement

The CIoU was prone to the imbalance phenomenology of sample quality, hence reducing the model generalization capability during training. In this paper, the dynamic focusing mechanism was employed to measure the loss between the predicted box and the truth. The anchor box quality was evaluated by an outlier measured and constructed distance attention, rather than the aspect ratio parameterized IoU. So the penalties for the high-quality anchor boxes can be reduced, while the harmful gradients from low-quality ones can be mitigated. The overall detection performance for UAVs can be improved accordingly.

The base version of FIoU was expressed as

$$\ell_{\mathrm{FIoU}_b} = \exp\left(\frac{(x - x^{gt})^2 + (y - y^{gt})^2}{((w^c)^2 + (h^c)^2)} \right) \cdot \ell_{\mathrm{IoU}} \tag{10}$$

where ℓ_{IoU} was the original intersection over union. To leverage the advantages of a non-monotonic focusing mechanism, the anchor box quality was reflected through mutual comparison. The base FIoU was then parameterized by the constants α, the outlier measure β, and the gradient gain γ, forming the advanced FIoU loss function:

$$\ell_{\mathrm{FIoU_a}} = \gamma \cdot \ell_{\mathrm{FIoU_b}} = \frac{\beta}{\delta \alpha^{(\beta - \delta)}} \cdot \ell_{\mathrm{FIoU_b}} \tag{11}$$

where γ, β, δ are the hyper-parameters.

As shown in Figure 10, This balanced gradient information from anchor boxes of different qualities during training effectively addressed the challenge of detecting small targets.

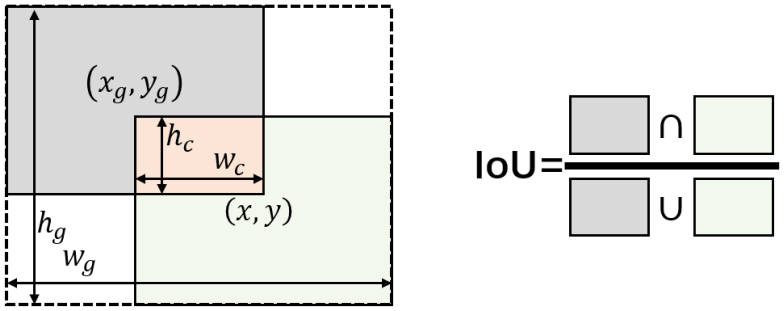


Figure 10. The illustration of dynamic focusing loss measurement.

4. Experiments and Discussion

In this section, the proposed method is verified. A series of comparative experiments were conducted. The experimental setting and the evaluation measurements are first described, followed by the ablation and the comparison with the state of the art. The results demonstrated that the detection capability for small-target drones in complex optical scenarios was improved by the proposed method. It is therefore suitable for the practical requirements of counter-drone systems.

4.1. Experimental Settings

4.1.1. The Dataset

The dataset was constructed using visible-light drone video footage captured by a multispectral sensor (manufactured in Changsha, China). The images were extracted from each frame, followed by removal of pure background and irrelevant frames.

To make the detection task more challenging, some issues were considered in the dataset collection.

- Various kinds of weather conditions were covered in the self-collected dataset, including sunny days, cloudy weather, dusk, and snowy conditions. So both the ideal weather and the adverse weather were considered in the self-collected dataset.
- The scenes near buildings and wooded areas were covered in the data collection. So both the normal conditions and the abnormal conditions were considered in the self-collected dataset.
- The publicly related video data (phantom02, phantom14, phantom22, phantom23, phantom43, phantom60) from the ARD-MAV dataset were incorporated. So multiple challenging scenarios, including complex backgrounds, non-planar scenes, target occlusion, motion blur, and small drone targets, were considered in the self-collected dataset.

4.1.2. The Parameters

The experiments were conducted on a server running Ubuntu 18.04, equipped with an Intel Xeon Gold 5120 CPU @ 2.2 GHz and an NVIDIA GeForce RTX 2080 Ti GPU (11 GB VRAM). The programming environment used Python 3.10, with the framework built on PyTorch 1.12.0. The settings for model training are delineated in Table 1.

It should be noted that all methods to be studied were retrained under identical conditions and with the same settings, so these methods can be compared fairly.

Table 1. The experimental settings.

Resolution	1920 × 1080	Optimizer	SGD
Target size	0.04%	learning rate	1×10^{-2}
Images	1500	Momentum	0.9
Training	900	Weight decay	5×10^{-4}
Testing	300	Batch size	16

4.1.3. Evaluation

Following the standard evaluation strategy, the UAV detection performance is measured by the precision rate (P), the recall rate (R), and the mean average precision (mAP). Let us simply review those measurements. The confusion matrix for target detection is shown in Table 2. It is composed of four situations: the true positive, the true negative, the false positive, and the false negative.

Table 2. The target detection measurements.

	Positive (T)	Negative (T)
Positive (P)	True Positive (TP)	False Positive (FP)
Negative (P)	False Negative (FN)	True Negative (TN)

- TP is an instance in which the detector predicts positive when the truth is indeed positive, i.e., a detection for which IoU is bigger than the threshold.
- FP is a wrong positive detection, i.e., a detection for which IoU is lower than the threshold.
- FN is an actual instance that is not detected.
- TN is a negative detection given a true negative instance.

The precision rate is the proportion of correct positive predictions:

$$\text{Precision}(\alpha) = \frac{TP}{TP + FP} = \frac{TP}{\text{Predictions}} \tag{12}$$

where α is the detection threshold. The recall rate is the proportion of true positives detected among all truths:

$$\text{Recall}(\alpha) = \frac{TP}{TP + FN} = \frac{TP}{\text{Truths}} \tag{13}$$

F1-Score is the harmonic mean of precision and recall:

$$\text{F1-Score}(\alpha) = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{14}$$

The mean average precision is the precision averaged across all values of recall between 0 and 1. It is derived from the Precision–Recall curve (PR):

$$\text{AP}(\alpha) = \int_0^1 P(\alpha) d\alpha \tag{15}$$

The precision at each recall bin α is interpolated by taking the maximum precision measured for a method for which the corresponding recall exceeds α:

$$P(\alpha) = \max_{\tilde{\alpha}:\tilde{\alpha} \geq \alpha} P(\tilde{\alpha})$$

4.2. The Verification of the Attention Mechanism

To verify the effect of the presented attention on UAV detection, two rounds of experiments were performed, in which the attentions were deployed on the backbone and the neck, respectively.

4.2.1. Attention-Induced Backbone

The initial experiment was conducted by integrating attention mechanisms into the original YOLO architecture. Several representative mechanisms were compared, including squeeze-and-excitation (SE) [14], convolutional block attention module (CBAM) [32], efficient channel attention (ECA) [25], and efficient multi-scale attention (EMA) [33]; these were incorporated into the feature enhancement of the detection architecture. This approach deliberately avoided direct placement into the backbone feature extraction network to prevent potential degradation of original weights and preservation of initially extracted features. The experimental results are presented in Table 3.

Table 3. The experiments of attention mechanism in backbone.

Attention	P(%)	R(%)	mAP@0.5(%)
Original	92.3	83.4	87.4
+SE [14]	88.9	87.0	89.6
+ECA [25]	92.0	83.4	88.5
+EMA [33]	92.4	84.3	87.4
+CBAM [32]	90.9	84.3	88.2
Ours	95.5	83.7	90.1

As shown in Table 3, all of the attention mechanisms introduced in the backbone could improve the detection performance. The results also corroborated that powerful representations can be promoted with the attention mechanisms, so the first core idea of this paper for UAV detection was reasonable. As for the detection performance, these attention mechanisms demonstrated different accuracies. Taking SE as an example, the precision rate was 88.7%, poorer than the baseline, while the recall rate (87.0%) was better than it. The other variants of attentions achieved similar performance. In summary, the precision rate and the mean average precision for he proposed method were much better than these competitors, while the recall rate for SE was the best.

4.2.2. Attention-Induced Neck

This paper formed multi-scale representations of UAVs with a hierarchical attention mechanism-induced backbone. So another question is whether the attention was effective in the neck. To verify the problem, another round of experiments was performed, where different kinds of attention mechanisms were deployed on the neck module. The experimental results are presented in Table 4.

The detection results shown in Table 4 were significantly different from the ones in Table 3. As can be seen, the integration of an attention mechanism into the feature enhancement network (neck) does not improve the detection performance. All of these attention-induced necks degraded the overall metric of mAP@0.5. Taking the ECA-enhanced neck as an example, the precision rate even dropped from 95.1% to 90.1%. Similarly, the mAP@0.5 for the CBAM-enhanced neck was 1.5% poorer than the standard configuration. The results also proved that the attention mechanisms may not always be effective during feature learning. It is therefore reasonable to keep the configuration of TSF unchanged.

Table 4. The experiments of attention mechanism in neck.

Attention	P(%)	R(%)	mAP@0.5(%)
TSF	95.1	83.1	89.0
+SE	90.8	86.4	88.5
+ECA	90.1	86.1	88.9
+EMA	91.2	84.0	87.6
+CBAM	93.0	83.7	87.5

4.3. The Verification of Loss Measurement

This paper presents the dynamic focusing loss measurement for UAV detection. To validate the performance of the presented method, a comparative experiment was conducted with the detection model. The presented loss measurement was compared with some commonly used loss metrics, including complete IoU (CIoU) [30], minimum points distance IoU (MPDIoU) [34], powerful IoU (PIoU) [35], and similarity weighted IoU (SIoU) loss functions [36]. The experimental results are shown in Table 5.

Table 5. Experimental performance of bounding box loss function.

LOSS	P(%)	R(%)	mAP@0.5(%)
CIoU [30]	90.1	86.1	87.9
PIoU [35]	90.3	81.3	87.1
MPDIoU [34]	87.1	81.0	81.8
SIoU [36]	89.2	86.7	88.1
Our	95.5	83.7	90.1

All of the loss measurements can be viewed as rooting from the original IoU, so the detection models with them can converge successfully during training. As for the detection performance, this is slightly different. For the metric of precision rate, the loss of the proposed measurement produced the best accuracy, 95.5%. It was 8.4% better than the competitor MPDIoU. Contrarily, the recall rate for SIoU was better than the remaining measurements. It was 5.7% better than MPDIoU. As for the overall detection performance, the proposed method was much better than these baseline measurements. It was 2.2%, 3.0%, 8.3%, and 2.0% better than CIoU, PIoU, MPDIoU, and SIOU. The results proved that the presented measurement was much more suitable for UAV detection in the complex scenarios.

4.4. The Comparison with the Baselines

In this paper, a new architecture composed of the attention-induced backbone and quasi-residual neck is presented for UAV detection. So it was important to compare the proposed method with the baselines, especially in complex environments. The richer features resulting from the high-quality baseline networks could promote inference precision, while lightweight networks could increase the inference speed, making them suitable for resource-constrained devices.

4.4.1. The Accuracy

In this section, the typical versions of YOLO, including YOLOv5, YOLOv6, YOLOv7 [37], YOLOv8, YOLOX and the two-stage detector (Faster-RCNN), were employed as baseline networks. The precision rate, the recall rate, and the mAP@0.5 were compared. The experimental results are shown in Table 6.

Table 6. The comparisons with the baselines.

Model	P(%)	R(%)	mAP@0.5(%)
Faster-RCNN	86.7	76.1	79.3
YOLOX-tiny	87.5	78.4	81.2
YOLOv5n	88.9	81.9	83.6
YOLOv6n	88.0	66.6	71.6
YOLOv7-tiny	92.3	83.4	87.4
YOLOv8n	91.0	65.4	75.5
Ours	95.5	83.7	90.1
$\uparrow_{min}$	3.2%	1.8%	2.7%
$\uparrow_{max}$	8.0%	18.3%	18.5%

In terms of competitive object detection, the family of YOLO algorithms demonstrated solid detection performance on the MAV dataset. However, the performance for UAV targets was much poorer, due to the limitations of the network architecture. As can be seen, the recall rate for YOLOv6n even dropped to 66.6%. The corresponding mAP@0.5 is 71.6%. Similar performance can be observed from the other algorithms. For the proposed method, the precision rate is 95.5%, 8.8% better than Faster-RCNN. The overall metric mAP@0.5 is 90.1%, 2.7% better than the nearest competitor, YOLOv7.

4.4.2. The Speed

In this paper, the configuration of the backbone and neck is improved by the attention mechanism and the quasi-residual connection, respectively, so it was necessary to verify whether the inference speed was impacted. To verify the question, another round of experiments was performed. We paid more attention to the inference efficiency and the speed. The experimental results are shown in Table 7, where the mAP and FPS (Frames Per Second) are listed.

Table 7. The trade-off between accuracy and speed.

Model	mAP@0.5(%)	FPS
Original	87.4	103.96
Ours	90.1	100.86

As can be seen, the computational cost of the proposed method has been slightly enhanced, because the attention mechanism and the quasi-residual connection are introduced in the backbone and neck. The inference speed only dropped from 103.96 FPS to 100.86 FPS. However, the detection accuracy has been improved from 87.4% to 90.1%. So the improvements of this paper are meaningful.

4.4.3. The Generalization Ability

To further test the generalization ability of the proposed method, the proposed method was verified with a publicly released dataset, Detfly. It is a dataset specifically designed for air-to-air visual unmanned aerial vehicle detection tasks. This dataset aims to solve the problem of detecting another flying drone from a mobile drone platform and is the first dataset in this field to systematically evaluate deep learning algorithms. The details on Detfly are shown in Table 8. The detection performance obtained by the proposed method is shown in Table 9.

As can be seen, the detection performance for the proposed method in the Detfly experiments is better than that in the self-collected dataset. The mAP is 2.6% better than the self-collected dataset. The results prove that the experimental setting in the self-collected

dataset was much more challenging than DetFly, even though the total number of images available was much less.

Table 8. The dataset of Detfly.

Resolution	Numbers	Background	Views	Challenges
4K	13,000	Sky City Field Mountain	Forward view Top-down view Bottom-up view	Strong/low light Motion blur Partial occlusion

Table 9. The verification of generalization with public dataset.

Dataset	Precision	Recall	mAP
DetFly	0.913	0.934	0.927

4.5. Visualization of the Detection Results

4.5.1. The Typical Scenes

The proposed method borrowed ideas from the classical YOLO algorithm, and was specific to the task of UAV detection in visible-light images. To verify the performance of the proposed method in visual detection, several challenging scenarios with complex backgrounds were presented. The results are shown in Figure 11.

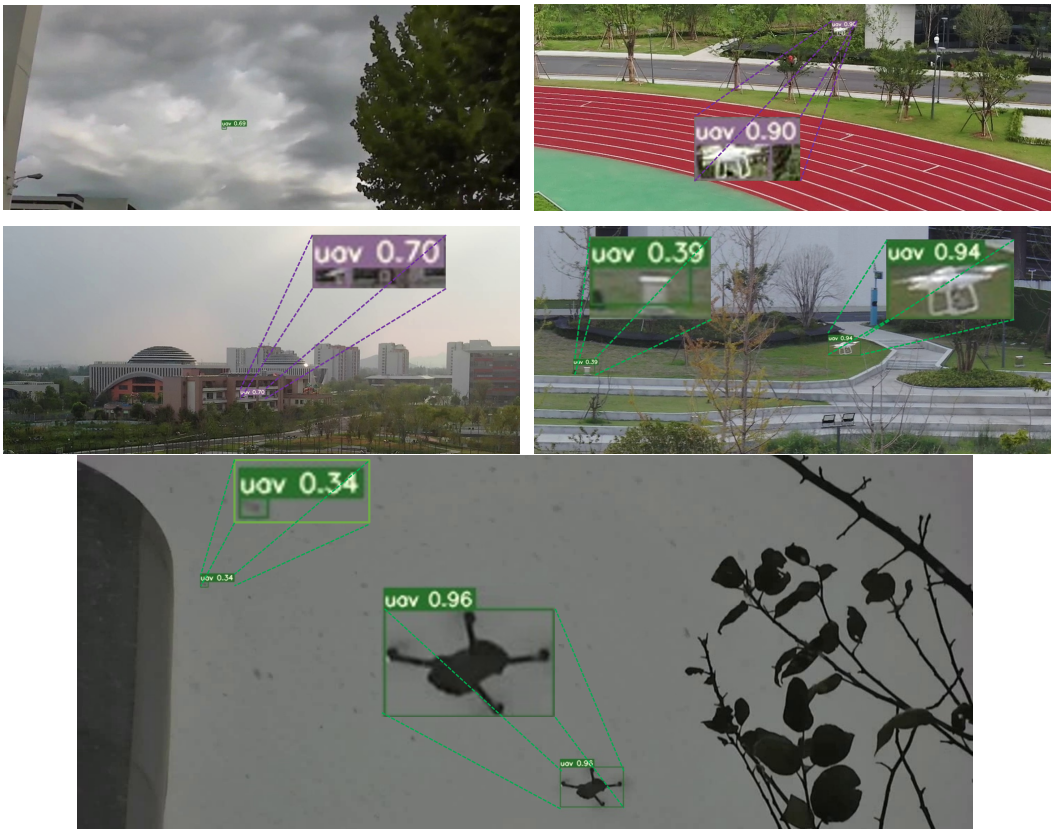


Figure 11. The detection results of proposed method.

Four kinds of typical scenes are considered. In the first scene, the drone is in a cloudy and low-light environment. The main challenges are the motion blur of the drone, as well as the poor lighting conditions. The features extracted by the general algorithms have limited representation power, and hence, it is difficult to locate the object accurately. The shortages can be improved by the proposed method, while the multi-scale representations of

the drone were obtained by the hierarchical attention mechanisms, so the drone can be located accurately. In the second scene, the UAV in the vegetation occlusion environment is considered. There are many trees at different heights, so the UAVs can be shielded easily. Even in adverse conditions, the proposed method always recognized the drones accurately. In the third and the fourth scenes, UAVs in buildings and dense vegetation background environments are considered. As can be seen, it is very difficult to find the drones in visual images. The proposed method could suppress the interference factors in the background effectively. The reliable detection capability was maintained even for very small targets. In summary, even under challenging conditions like complex backgrounds, low light, and motion blur, the proposed method still demonstrated robust and accurate detection precision.

4.5.2. The Feature Heatmap

To further visualize the learned feature, Figure 12 displays the heatmap of Grad-CAM (Gradient-weighted Class Activation Mapping) generated by the proposed method. It is a popularly used trick to visualize the features learned by the network. Grad-CAM paid more attention to the key regions.

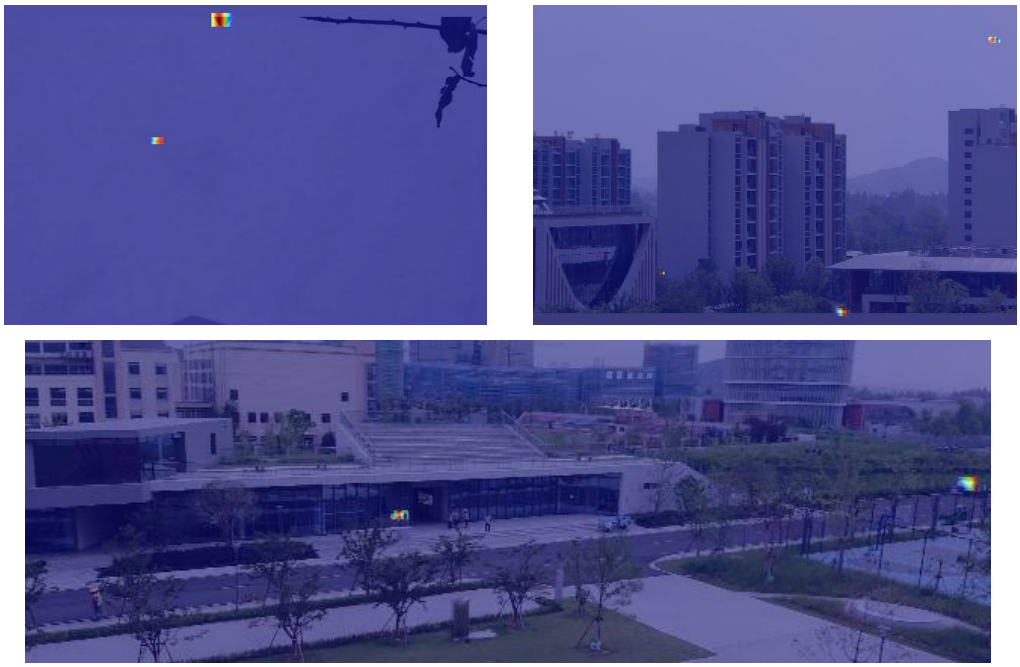


Figure 12. The heatmap generated by proposed method.

As can be seen, the results demonstrated that the detection accuracy, the localization capability, and the boundary delineation produced by the proposed method were much better. It was capable of producing stronger robustness and higher precision, even in complex scenes and for the small targets. These results also revealed the distinct patterns in feature learning and regional attention between the models.

5. Conclusions

In this paper, a hierarchical attention-induced cross-scale learning framework was developed for UAV detection in visible images. The multi-scale representations of the target were first produced by the hierarchical attention mechanism. They were further used to build the quasi-residual connected neck. The dynamic focusing loss measurement was presented to promote the learning efficiency. Multiple comparative experiments are performed, from which some conclusions can be obtained.

- It is reasonable to locate and detect the drones by building the multi-scale representations of target.
- The structure of the quasi-residual connected neck was very suitable for the task of UAV detection.
- Drone detection, even in adverse lighting and weather conditions, can be achieved by the proposed method.

In the future, two kinds of problems will be focused on. First, more challenging conditions, such as adverse weather, motion blue, and hazing conditions, will be considered. Second, the integration of image quality improvement and target detection will be unified into a single framework.

Author Contributions: Conceptualization, W.W. and X.Y.; experiments, Y.Z. (Yan Zhang) and Y.Z. (Yaxiu Zhang); methodology, W.W.; validation, L.Z.; writing, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The self-constructed dataset presented in this study is not publicly available due to institutional confidentiality and security restrictions mandated by applicable laws and regulations. Access to this dataset is restricted to authorized personnel only. The publicly available ARD-MAV dataset used in this study has been properly cited in the manuscript (see Section 4.1.1). Further inquiries regarding the self-constructed dataset may be directed to the corresponding author, and any request will be considered in accordance with institutional data access policies.

Acknowledgments: We would like to thank the handling Associate Editor and the anonymous reviewers for the great contributions to this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lowe, D. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
2. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.
3. Strickland, R.; Hahn, H.I. Wavelet transform methods for object detection and recovery. *IEEE Trans. Image Process.* **1997**, *6*, 724–735. [CrossRef] [PubMed]
4. Bay, H.; Ess, A.; Tuytelaars, T.; Gool, L.V. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
5. Arandjelovic, R.; Zisserman, A. Three things everyone should know to improve object retrieval. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2012; pp. 2911–2918.
6. Dong, J.; Soatto, S. Domain-size pooling in local descriptors: DSP-SIFT. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2015; pp. 5097–5106.
7. Sun, Y.; Sun, H.; Yan, L.; Fan, S.; Chen, R. RBA: Reduced bundle adjustment for oblique aerial photogrammetry. *ISPRS J. Photogramm. Remote Sens.* **2016**, *121*, 128–142. [CrossRef]
8. Sedaghat, A.; Ebadi, H. Remote Sensing Image Matching Based on Adaptive Binning SIFT Descriptor. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 5283–5293. [CrossRef]
9. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D. Cascade object detection with deformable part models. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2010; pp. 2241–2248.
10. Wu, X.; Zhang, H.; Zhao, C.; Li, Z.; Wang, Y. Improved YOLOv4 for real-time detection algorithm of low-slow-small unmanned aerial vehicles. *J. Appl. Opt.* **2024**, *45*, 79–88.
11. Hao, J.; Zhang, J.; Zhang, S.; Zhang, Z.; Hao, Z.; Dai, S.; Ran, N. UAV intrusion detection method based on improved YOLO. *J. Electron. Meas. Instrum.* **2024**, *38*, 143–151.
12. Wei, F.; Zhou, J.P.; Tan, X.; Lin, J.; Tian, L.; Wang, H. Lightweight YOLOv5 detection algorithm for low-altitude micro UAV. *J. Optoelectron. Laser* **2024**, *35*, 641–649.
13. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 116–131.

14. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
15. Cheng, Q.; Wang, Y.; He, W.; Bai, Y. Lightweight air-to-air unmanned aerial vehicle target detection model. *Sci. Rep.* **2024**, *14*, 2609. [PubMed]
16. Yuan, Y.; Gao, S.; Zhang, Z.; Wang, W.; Xu, Z.; Liu, Z. Edge-Cloud Collaborative UAV Object Detection: Edge-Embedded Lightweight Algorithm Design and Task Offloading Using Fuzzy Neural Network. *IEEE Trans. Cloud Comput.* **2024**, *12*, 306–318.
17. Ye, T.; Qin, W.; Zhao, Z.; Gao, X.; Deng, X.; Ouyang, Y. Real-Time Object Detection Network in UAV-Vision Based on CNN and Transformer. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2505713.
18. Jiao, Z.; Wang, M.; Qiao, S.; Zhang, Y.; Huang, Z. Transformer-Based Object Detection in Low-Altitude Maritime UAV Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4210413.
19. Zou, Z.; Hu, X.; Zhong, P. Active Object Detection for UAV Remote Sensing via Behavior Cloning and Enhanced Q-Network with Shallow Features. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5628116. [CrossRef]
20. Zhang, T.; Zhang, X.; Gao, G. Density knowledge mining for quantity-aware marine vessel surveillance using satellite SAR data. *IEEE Trans. Ind. Inform.* **2026**, *22*, 5173–5184. [CrossRef]
21. Zhang, T.; Zhang, X.; Gao, G. TLSA: A Triple-Level Speckle Awareness Framework for Speckle-Robust Marine Ship Surveillance Using Satellite Synthetic Aperture Radar. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 1–14. [CrossRef]
22. Li, H.L.; Chen, S.W. General Polarimetric Correlation Pattern: A Visualization and Characterization Tool for Target Joint-Domain Scattering Mechanisms Investigation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *64*, 5200417. [CrossRef]
23. Li, H.L.; Chen, S.W. Polyhedral Corner Reflectors Multidomain Joint Characterization with Fully Polarimetric Radar. *IEEE Trans. Antennas Propag.* **2025**, *73*, 10679–10693. [CrossRef]
24. Dai, L.Y.; Chen, S.W. Context2Context: A Zero-Shot SAR Image Speckle Filter. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5204911. [CrossRef]
25. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020; pp. 11534–11542.
26. Du, X.; Zou, Y.; Lei, T.; Gu, D.; Zhang, X.; Nandi, A.K. MSD-EMA: Multiscale Decoupled Expectation–Maximization Attention for Polyp Segmentation. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5013113.
27. Yue, G.; Li, S.; Cong, R.; Zhou, T.; Lei, B.; Wang, T. Attention-Guided Pyramid Context Network for Polyp Segmentation in Colonoscopy Images. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5008213. [CrossRef]
28. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Carafe: Content-aware reassembly of features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019; pp. 3007–3016.
29. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [CrossRef]
30. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence (AAAI): Washington, DC, USA, 2020; Volume 34, pp. 12993–13000.
31. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
32. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 3–19.
33. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In *Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
34. Ma, S.; Xu, Y. MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. *arXiv* **2023**, arXiv:2307.07662.
35. Liu, C.; Wang, K.; Li, Q.; Zhao, F.; Zhao, K.; Ma, H. Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism. *Neural Netw.* **2024**, *170*, 276–284. [PubMed]
36. Kumar, U.R.; Vandewalle, P. Similarity-Weighted IoU (sIOU): A Comprehensive Metric for Evaluating Model Performance Through Similarity-Weighted Class Overlaps. In *Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2024; pp. 936–942.
37. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 7464–7475.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.