DASSF: Dynamic-Attention Scale-Sequence Fusion for Aerial Object Detection

Haodong Li and Haicheng Qu*
Liaoning Technical University, School of Software
Huludao, 125105, China
472321734@stu.lntu.edu.cn and quhaicheng@lntu.edu.cn

Abstract

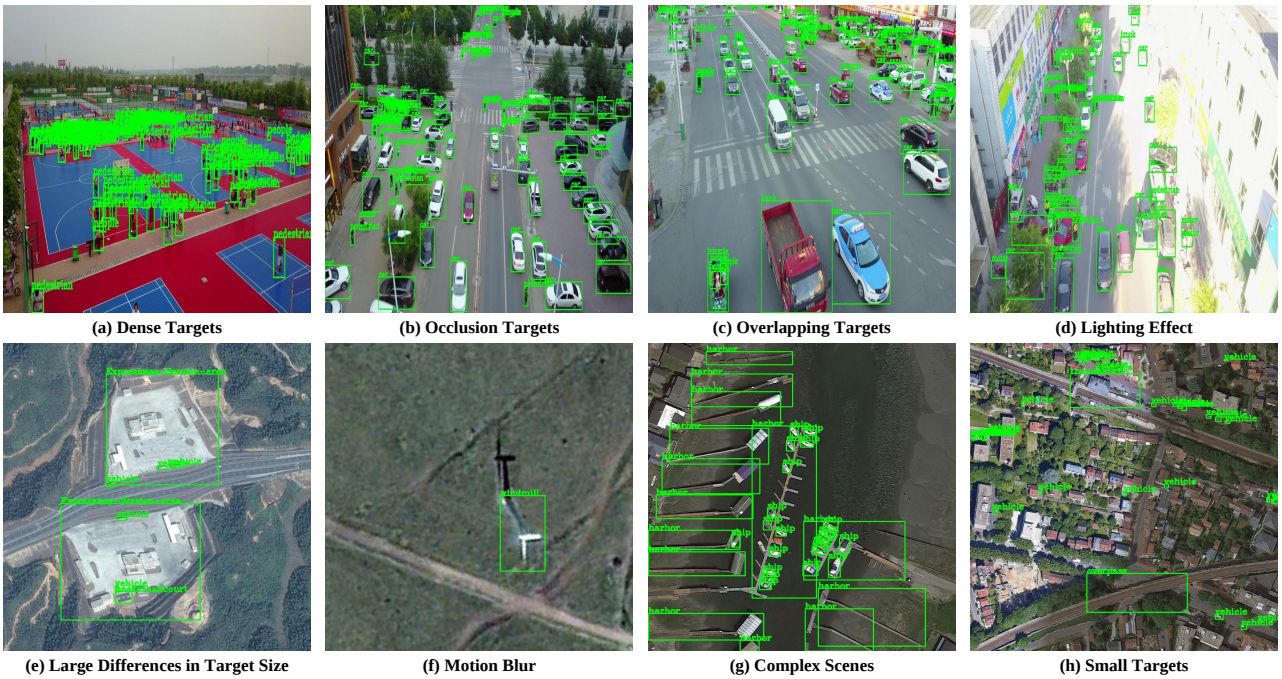
The detection of small objects in aerial images is a fundamental task in the field of computer vision. Moving objects in aerial photography have problems such as different shapes and sizes, dense overlap, occlusion by the background, and object blur, however, the original YOLO method has low overall detection accuracy due to its weak ability to perceive targets of different scales. In order to improve the detection accuracy of densely overlapping small targets and fuzzy targets, this paper proposes a dynamic-attention scale-sequence fusion method (DASSF) for small target detection in aerial images. First, we propose a dynamic scale sequence feature fusion (DSSFF) module that improves the upsampling mechanism and reduces computational load. Secondly, a x-small object detection head is specially added to enhance the detection capability of small targets. Finally, in order to improve the expressive ability of targets of different types and sizes, we use the dynamic head (DyHead). The model we proposed solves the problem of small target detection in aerial images and can be applied to multiple different versions of the YOLO method, which is universal. Experimental results demonstrate that when the DASSF method is applied to YOLOv8, it achieves a 10.2% and 4.2% improvement in mean Average Precision (mAP) on the VisDrone-2019 and DIOR datasets, respectively, compared to YOLOv8n. This performance surpasses that of current mainstream methods. Additionally, when the DASSF method is integrated into different versions of the YOLO model, the detection performance for aerial images significantly improves compared to the baseline models.

Keywords: Aerial images, Small target detection, Feature fusion, Upsampling method.

1. Introduction

Object detection is now widely applied in fields such as intelligent transportation [6], medical diagnosis [20], industrial manufacturing [35], and re-identification [23]. With the ongoing advancements in drone technology and the growing maturity of remote sensing technology, aerial image object detection has emerged as a significant research area due to its immense potential. However, compared to traditional object detection in natural scenes, aerial images present unique challenges. These include a wide range of target scales, small object sizes, diverse angle variations, complex backgrounds, and vulnerability to solar radiation and atmospheric conditions, as illustrated in fig. 1. These factors greatly complicate the accurate detection and recognition of small targets in aerial images. The YOLO series networks, while popular, lack effective feature fusion and scale perception capabilities for objects of varying sizes and complex shapes in aerial images. In contrast, two-stage convolutional neural network (CNN) detection methods or Transformer-based DETR series methods consume excessive computational resources. Therefore, further research and innovation are essential to improve both the accuracy and efficiency of object detection in aerial images under these conditions.

According to the definition in the MS COCO dataset [14] for object detection, small targets refer to objects with a size of less than 32×32 pixels. Detecting small targets is crucial in many real-world applications. For instance, in agricultural pest control, small target detection technology can be used to identify pests in crops. By accurately detecting these tiny targets, it enables farmers to take timely prevention and control measures, ensuring the healthy growth of crops. In medical image analysis, small target detection is widely applied in cell detection and lesion identification. This is especially important in the early diagnosis of tumors, where accurate identification of tiny lesions is critical for improving treatment outcomes and increasing patient survival rates. Similarly, in marine biological monitoring, small target detection technology is employed to identify microorganisms and plankton communities in the ocean, providing valuable insights for ecological research and environmental protection.While existing object detection methods have made progress in these fields, small targets often occupy very small pixel areas in images and can easily blend into complex backgrounds, making it difficult



(a) Dense Targets (b) Occlusion Targets (c) Overlapping Targets (d) Lighting Effect

(e) Large Differences in Target Size (f) Motion Blur (g) Complex Scenes (h) Small Targets

Figure 1. The primary challenges and obstacles encountered in aerial image object detection.

for traditional methods to effectively extract fine features. This results in low detection accuracy. Additionally, factors such as object overlap and image blur further complicate small target detection, limiting the performance of current methods when dealing with multi-scale targets.

In this paper, to enhance the detection performance of the YOLO series network model in aerial images, particularly for small targets, we propose a YOLO model integrated with Dynamic-Attention Scale-Sequence Fusion (DASSF). The main contributions of this paper can be summarized as follows:

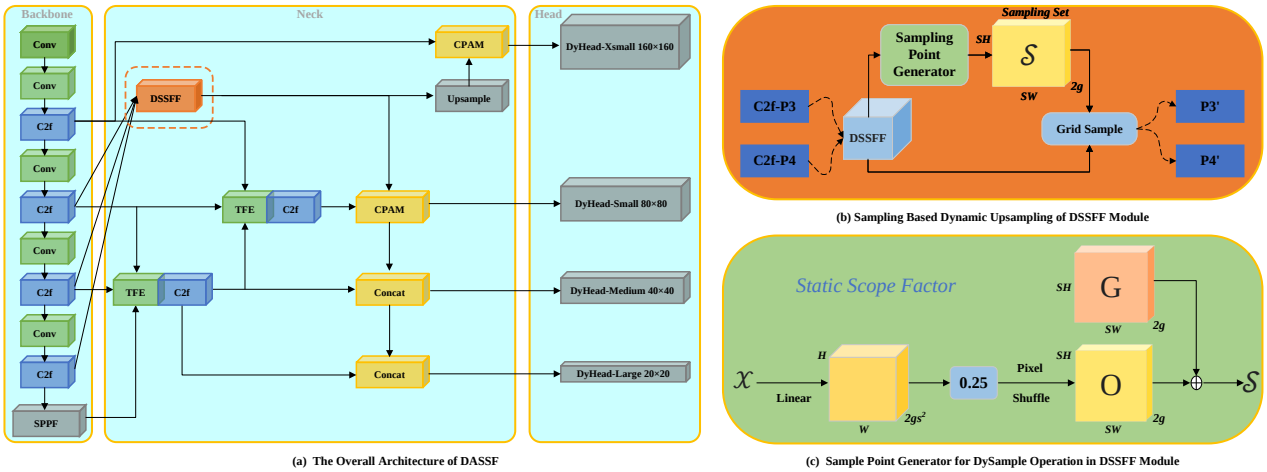
- We propose a new and effective method, DASSF, for aerial image object detection tasks, which combines multiple modules to effectively improve the detection of objects across different scales and categories, especially small objects.
- We design a dynamic scale sequence feature fusion (DSSFF) module to accurately and efficiently extract global high-level semantic information in images of different scales, which not only reduces the model complexity but also can accurately detect small objects through point sampling.
- We conduct detailed and comprehensive comparison and ablation experiments on the proposed DASSF method, which demonstrates its effectiveness. Additionally, we combine it flexibly with multiple YOLO series methods to showcase the versatility of the approach.

2. Related Work

2.1. Aerial Image Object Detection

In recent years, research on aerial image target detection has mainly focused on improving model performance by improving feature extraction and context learning. For example, CFIL [29] introduces a frequency domain feature extraction module [30] and a frequency domain feature interaction mechanism to enhance the ability to extract significant features and better distinguish targets in complex backgrounds. MFC [21] proposes a frequency domain filtering module to further enhance the feature expression ability of dense targets, thereby performing well in processing complex scenes. LR-FPN [12] improves remote sensing target detection by strengthening low-level position information and fine-grained context interaction, especially in application scenarios such as agriculture and urban planning. In addition, PBSL [22] can highlight relevant target features in aerial images and suppress irrelevant information by introducing a multimodal alignment method, thereby improving the robustness of overall detection.

Although these methods have achieved good results in aerial image target detection tasks, especially in the detection of dense targets and overlapping targets, they still have some limitations. First, most methods mainly improve the



(a) The Overall Architecture of DASSF

(b) Sampling Based Dynamic Upsampling of DSSFF Module

(c) Sample Point Generator for DySample Operation in DSSFF Module

Figure 2. (a) The overall architecture of our proposed DASSF. DSSFF refers to the dynamic scale sequence feature fusion module we propose. (b) and (c) show schematic diagrams of the DSSFF module, DySample. S represents the upsampling ratio, G denotes the grid sampling point coordinates, and O indicates the point position offset generated by the dynamic sampling point generator. SH and SW represent the sampling height and width, respectively. gs^2 refers to the number of channels in the feature map after the linear layer. The remaining components are: TFE, the triple feature encoding module; CPAM, the channel and position attention mechanism; and DyHead, the dynamic detection head.

average detection accuracy of the overall target, but the detection effect of small targets is limited, especially under complex backgrounds and blurred conditions, the detection accuracy is still not ideal. Secondly, many existing methods are often optimized for specific model structures, lack versatility, and are difficult to flexibly apply under different detection tasks or model frameworks, which limits their widespread promotion in practical applications.

2.2. Small Object Detection

As a major challenge in the field of target detection, small target detection has received extensive attention in recent years. Existing research mainly improves the detection performance of small targets by integrating attention mechanisms, optimizing loss functions, and extracting features in stages. For example, the InsDist [10] method combines feature-based and relationship-based knowledge distillation to improve the detection effect of small objects in remote sensing images; in the field of agricultural control, GA-SGD [32] guides the model to focus on the detection of small pests by designing selection, crossover, and mutation operations, thereby improving the detection effect in complex environments; in addition, LESPS [33] uses a point supervision method for the detection of infrared small targets, which greatly reduces the cost of manual annotation and improves detection efficiency; UBDDM [17] achieves multi-scale recognition of significant areas in the image by constructing a small target perception module, and achieves good results in the two-stage bolt defect detection task.

Although these methods have made significant progress in small target detection tasks in different application fields, they still have some shortcomings. First, although the two-stage detection method improves the detection accuracy, its complex architecture greatly increases the computational overhead and inference time of the model, making it difficult to achieve efficient operation in application scenarios with limited resources or high real-time requirements. Secondly, the knowledge distillation method enhances the model's perception of small targets through a large amount of training, but the distillation process usually consumes a lot of computing resources, and the training process is complex and time-consuming, which is difficult to meet the needs of large-scale practical applications. Finally, many existing methods still have the problem of insufficient accuracy when dealing with dense and overlapping small targets, especially when the background is complex or the target is blurred, the detection performance is relatively limited. These problems limit the performance of existing small target detection methods in efficient and accurate detection.

3. Method

3.1. The Overall Architecture of DASSF-YOLO

Fig.2 shows the overall architecture of the dynamic-attention scale-sequence fusion method (DASSF) network model we designed. We use CSPDarknet53 as the backbone network. We utilize ASF-YOLO [9]'s neck network to enhance the detection of small, dense, and blurry targets in aerial images. The neck of the network utilizes the TFE in ASF-YOLO twice for fusing feature maps of different di-

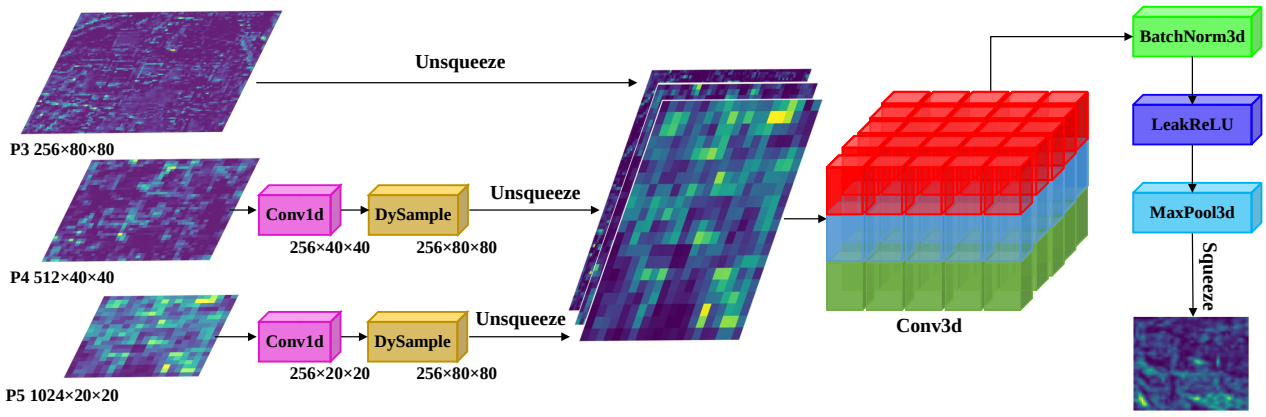


Figure 3. The structure of the DSSFF. The features are extracted efficiently and accurately through dynamic upsampling, feature map stacking, and 3D convolution normalization activation operations. The detailed process of dynamic upsampling is shown in algorithm 1.

mensions to obtain rich global channel information. Then apply the DSSFF to obtain accurate local location information. Then, the processed information is effectively interacted through the CPAM. Finally, by adding the x-small detection head and applying the dynamic head (DyHead). The model can detect objects of various scales in aerial images.

3.2. Improvements to the Neck

3.2.1 Triple Feature Encoding Module

The TFE module is a feature fusion mechanism. First, adjust the number of channels of the large, medium and small feature layers to make them equal through CBS operation. Then, the large-scale feature map is subjected to a downsampling operation of maximum pooling + average pooling, which helps to retain the high-resolution features and the diversity of semantic information of different objects in aerial images; for small-scale feature maps, the nearest neighbor interpolation method is used for upsampling, which can maintain the richness of local features of low-resolution images and prevent the loss of small target location feature information. Finally, feature maps of different scales are fused through concat operations.

3.2.2 Dynamic Scale Sequence Feature Fusion Module

The original SSFF module was designed for the P3 layer and is a key component used to process multi-scale information and has the ability to extract features of different scales. Scale means detail in the image. A blurry image may lose details, but the structural features of the image can be preserved, helping to solve the image blur problem in satellite remote sensing images. The input image of SSFF is in formula 1.

$$F_o(w,h) = G_o(w,h) \times f(w,h). \tag{1}$$

Where f(w, h) represents a 2D input image with width w and height h. $G_o(w, h)$ is the filter used for smooth convolution. This module contains two upsampling operations for the P4 and P5 layers. The nearest upsampling method in the original module will lead to the loss of key image details and requires a lot of calculation and parameter overhead. Therefore, we introduce DySample [16], an ultra-lightweight and effective dynamic upsampler to replace the nearest upsampling method in the original module. The structure is shown in (b) and (c) of fig.2.

We adopt a static factor sampling method, based on the theory of adjusting point sampling position offset, to dynamically create a sampling set of point positions through the sample point generator in the feature map. If the input feature map is X (with size $C \times H \times W$), the offset O is obtained by the network through projection and multiplication by the static factor g.

$$O = g \times \text{Linear}(X). \tag{2}$$

Among them, Linear is a linear layer used to generate the offset O based on the input feature X, and g is a constant that controls the size of the offset.

$$S = G + O. \tag{3}$$

$$X' = \text{Grid\_sample}(X, S). \tag{4}$$

The generated offset O is added to the original sampling grid G and then passed into the grid sampling function along with the original feature map X for upsampling. This sampling point offset upsampling mechanism helps feature

Algorithm 1 The process of dynamic upsampling

Input: The feature map x of size C × H × W, upsampling multiple scale and groups default to 4.

Output: The feature map X' of size C × scale × H × scale × W.

1: Generate a 2D convolutional layer for offset and perform normal distribution initialization, and get out_1.
2: Generate the initial position for offset calculation, and get out_2.
3: Apply out_1 to the input x and adjust the range by × 0.25, then append the offset of out_2, and then transform it into 2 × -1 × H × W, and get out_3.
4: Create a normalized target coordinate grid, and get out_4.
5: Add out_3 and out_4 and normalize them to the [-1,1] interval, and get out_5.
6: Use pixel_shuffle to upsample coordinates for out_5 and adjust the output size, then use grid_sample for bilinear interpolation sampling, and get out.
7: **return** out.

maps of different scales better distinguish boundary areas and extract object information at varying scales. Moreover, by controlling the offset position, the overlap of sampling points is reduced, the positions of sampling points are optimized, and thus the computational complexity is minimized. The structure of the DSSFF module is shown in fig. 3. The pseudocode for dynamic upsampling with a static sampling factor of 0.25 is shown in algorithm 1.

3.2.3 Channel and Position Attention Mechanism

The CPAM integrates the DSSFF and TFE modules, which focus on information-rich channels and small object features related to spatial location. This allows the model to more accurately identify and locate small targets in images, thereby improving the detection capabilities of detailed small objects. Input 1 is the detailed features after TFE processing as channel attention network, which is used to effectively capture cross-channel interactions. This is an attention mechanism that does not require dimensionality reduction. The capture of local cross-channel interactions is achieved using 1D convolutions of size k, where the kernel size k represents the coverage of local cross-channel interactions. Using the output of the channel attention mechanism and the feature map processed by DSSFF as the input of the position attention network, the position information of different targets can be extracted.

3.3. Improvements to the Head

3.3.1 Dynamic Head

The aerial imagery targets used in this study presented complex backgrounds. Because the target in the drone image is blocked by houses and trees and the target in the remote sensing image is affected by light and clouds. The scale of the target is easy to change, and the image is easy to become blurred and distorted. Therefore, it is crucial that the detection method has a full range of perception capabilities. DyHead [3] was proposed by Dai et al, which simultaneously combines scale-aware attention (π_L) in formula 5, spatial-aware attention (π_S) in formula 6, and task-aware attention (π_C) in formula 7, enhances the model's adaptability to various target sizes, understanding of object placement, and context understanding.

$$\pi_L(F) \cdot F = \sigma\left(f\left(\frac{1}{S \cdot C}\sum_{S,C} F\right)\right) \cdot F. \tag{5}$$

Scale-aware attention (π_L) performs average pooling on the input feature map, then uses a 1 × 1 convolution layer and ReLU activation function for feature extraction, then uses the hard-sigmoid function to balance model accuracy and speed, finally the elements are multiplied with the input feature map.

$$\pi_S(F) \cdot F = \frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{K} w_{l,j} \cdot F\left(l; p_j + \Delta p_j; c\right) \cdot \Delta m_j. \tag{6}$$

Spatial-aware attention (π_S) first processes the input tensor using a 3 × 3 convolutional layer to obtain the offset value of the feature map and the weight term of the feature map offset, and then weights and sums all features.

$$\pi_C(F) \cdot F = \max\left(\alpha^1(F) \cdot F_C + \beta^1(F), \alpha^2(F) \cdot F_C + \beta^2(F)\right). \tag{7}$$

Task-aware attention (π_C) is first average pooling in the L × S dimension to reduce the number of channels. Subsequently, two fully connected layers are adopted and activated using the ReLU function and then passed through a normalization layer. Finally, different channel values are output according to different tasks to complete the task perception of the feature map.

4. Experiment and Analysis

In order to validate the superiority of the proposed DASSF method, we combine it with YOLOv8n and conduct comparison, ablation and general experiments on two

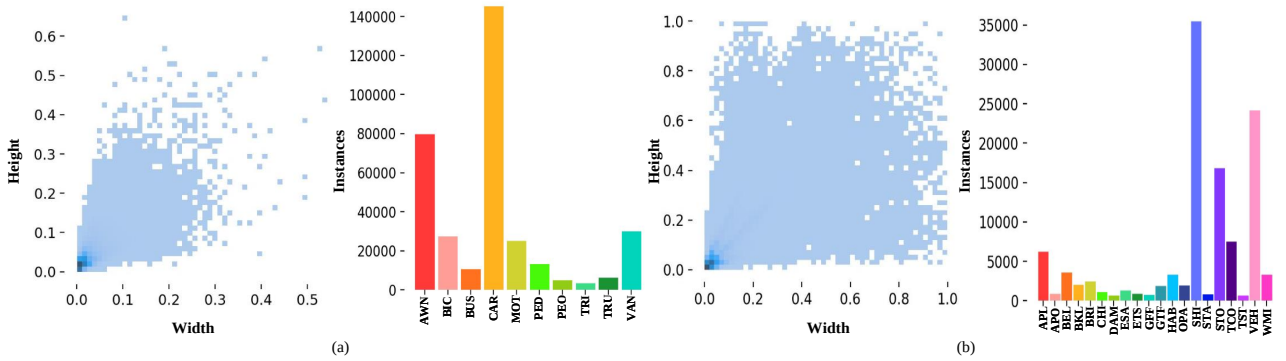


Figure 4. (a) Size distribution and object count of the VisDrone-2019 dataset; (b) Size distribution and object count of the DIOR dataset.

Table 1. Comparison of detection results of mainstream methods on the VisDrone-2019 dataset.

Method	Year	*Awn*	*Bic*	*Bus*	*Car*	*Mot*	*Ped*	*Peo*	*Tri*	*Tru*	*Van*	mAP50	mAP50:95
ATSS [34]	2020	10.8	14.1	41.3	74.2	36.3	37.9	18.5	20.5	32.6	36.2	32.2	19.8
Deformable-DETR [36]	2020	13.1	12.0	56.6	69.2	28.1	27.8	16.4	16.1	40.1	36.5	31.6	17.3
Conditional-DETR [19]	2021	8.2	11.3	33.5	62.8	36.5	30.4	25.6	20.2	30.5	30.2	28.9	15.0
DDOD [2]	2021	14.2	18.2	58.5	78.8	47.5	47.4	34.9	27.5	**41.0**	45.4	40.7	24.8
TOOD [5]	2021	14.2	**19.8**	56.4	79.3	49.2	46.8	35.4	27.2	40.9	45.6	38.8	24.3
Dab-DETR [15]	2022	15.3	12.7	57.5	66.7	26.2	21.4	14.3	19.7	39.3	38.2	31.1	15.5
YOLOv6n [11]	2022	11.1	5.0	42.9	73.5	31.1	30.0	24.6	18.0	24.2	35.3	29.6	17.1
DAMO-YOLO [31]	2022	11.8	7.2	42.6	75.1	35.6	34.2	27.4	21.2	27.3	36.9	31.9	18.3
RTMDET [18]	2022	14.6	12.1	56.1	75.2	40.4	34.2	28.3	25.1	36.3	41.4	36.4	21.5
YOLO-MS [1]	2023	10.9	7.4	44.7	74.5	34.1	32.8	26.2	19.7	24.3	36.9	31.1	17.6
Gold-YOLO [25]	2023	12.1	8.5	48.8	75.7	37.0	33.9	27.4	22.1	28.2	38.8	33.2	19.3
ASF-YOLO [9]	2023	11.0	7.2	43.8	74.4	33.4	32.8	25.8	19.1	26.6	37.2	31.1	17.9
Baseline [8]	2023	11.0	7.0	44.8	75.1	35.1	33.6	26.9	20.4	27.4	37.5	31.9	18.2
Ours	-	**17.5**	15.2	**57.6**	**82.9**	**49.7**	**47.6**	**39.3**	**27.8**	36.0	**47.2**	**42.1**	**25.2**

datasets. Throughout the experiments, we maintain consistency in hyperparameters and other experimental details. The comparison results show that our proposed method significantly improves the accuracy of small targets in aerial images. The comparison results show that our proposed method significantly improves the accuracy of small targets in aerial images.

4.1. Datasets

In our experiments, we use two datasets. The first is the Tianjin University AISKYEYE team publicly released the VisDrone-2019 [4] dataset. This dataset is designed for target detection in UAV images of remote sensing scenes with high diversity. The images are annotated with labels for ten categories, including awning-tricycle (Awn), bicycle (Bic), bus (Bus), car (Car), motorcycle (Mot), pedestrian (Ped), person (Peo), tricycle (Tri), truck (Tru) and van (Van). The dataset is divided into three distinct subsets: 6471 images for training, 548 images for validation, and 1610 images for testing. The second is the DIOR [13] remote sensing dataset was released by Northwestern Polytechnical University in 2018. This benchmark dataset contains 23,463 images and 192,472 instances for object detection in optical remote sensing images. It covers 20 common places and object categories, including airplane (AE), airport (AT), baseballfield (BD), basketballcourt (BT), bridge (BE), chimney (CY), dam (DM), Expressway-Service-area (EA), Expressway-toll-station (EN), golffield (GO), groundtrackfield (GR), harbor (HR), overpass (OS), ship (SP), stadium (SM), storagetank (SK), tenniscourt (TT), trainstation (TN), vehicle (VE), windmill (WL). The dataset is divided into three different subsets: 14077 images for training, 4694 images for validation, and 4692 images for testing. The size and category distributions of the two datasets are shown in fig. 4.

4.2. Implementation Details and Evaluation Metrics

The hardware configuration of this experiment includes: CPU: AMD EPYC 7551P 32-Core Processor, GPU: NVIDIA RTX A4000, Memory: 16G. The software environment includes: Ubuntu 20.04.1, python 3.8.10, Torch 1.13.1.

The hyperparameter settings are as follows: training is

Table 2. Statistics of all objects in the VisDrone-2019 dataset that are less than 32 × 32 pixels in size.

Category	Awn	Bic	Bus	Car	Mot	Ped	Peo	Tri	Tru	Van
Proportion (%)	58.5	82.8	47.4	51.0	83.2	89.0	92.0	61.4	40.5	49.3

conducted for 200 epochs, the batch size is set to 8, parameters are updated using stochastic gradient descent (SGD), the learning rate is set to 0.0001, the weight decay rate is 0.0005, and the intersection over union (IOU) threshold is set to 0.7.

The data augmentation settings are as follows: the image hue is set to 0.015, the image saturation is set to 0.7, the image brightness is set to 0.4, and the image flip probability is set to 0.5.

In order to evaluate the model's performance, this experiment uses precision (Pre), recall (Rec), mean average precision (mAP), and frames per second (FPS) as indicators.

4.3. Comparison with State-of-the-art Methods

4.3.1 Comparisons on VisDrone-2019

We use the VisDrone-2019 dataset to conduct comparative experiments with mainstream target detection methods, including ATSS, Defor-mable-DETR, Conditional-DETR, DDOD, TOOD, Dab-DETR, YOLOv6n, DAMO-YOLO, RTMDET, YOLO-MS, Gold-YOLO, ASF-YOLO. The comparative experimental results in table 1 show that our proposed method surpasses the selected target detection method YOLOv8n and improves the detection accuracy by 10.2%. And it surpassed other mainstream target detection methods in the table, achieving new SOTA results of 42.1% and 25.2% in mAP50 and mAP50:95 respectively. The best detection accuracy was achieved in 8 out of 10 categories. Due to the addition of a small target detection head and the application of the DSSFF module, the model's ability to detect small targets and extract and fuse features of objects of different scales are enhanced. Excellent detection results can be achieved for large-sized targets such as buses and garbage trucks, as well as small-sized targets such as pedestrians and motorcycles.

Additionally, to more intuitively demonstrate the impact of our proposed DASSF method on small target detection in aerial images, we calculate the proportion of all objects in the VisDrone-2019 dataset that are smaller than 32 × 32 pixels. The statistical results are presented in table 2. Combined with our comparative experimental results in table 1, it is evident that the DASSF method excels in small target detection. With the exception of the 4th (Bic) and 10th (Tru) categories in the statistics for objects smaller than 32 × 32 pixels, all other object categories achieved the highest detection accuracy.

4.3.2 Comparisons on DIOR

Furthermore, we also compare the proposed method with mainstream methods on the DIOR dataset. As shown in table 3, the proposed method outperforms other mainstream target detection methods in terms of overall accuracy, achieving the best results in 12 out of 20 categories in the DIOR dataset, with an overall accuracy of 87.1%. Exceeding Conditional-DETR, TOOD, and RTMDET by 7.9%, 3.1%, and 1.8% respectively on mAP50. Due to the use of DyHead with self-attention, the model's ability to perceive objects of different scales in remote sensing images is enhanced. Moreover, the DSSFF module solves the problem of misdetection and leakage of targets such as chimneys and windmills that are affected by light, clouds and other factors. And due to the improved upsampling mechanism in the scale sequence feature fusion module, which reduces the amount of calculation, the FPS exceeds the ten target detection methods in the table, ensuring the real-time performance of the proposed method.

4.4. Ablation Studies and Analysis

4.4.1 Ablation experiments on different modules of DASSF method

We conduct ablation experiments on the proposed method on two datasets, and the experimental results are shown in table 4. Baseline is YOLOv8n. The detection accuracy index of the baseline model is at the lowest position. The mAP50 and mAP50:95 of the finally proposed improved model on the two datasets increased by 10.2%, 7.0%, 4.2% and 5.2% respectively compared with the baseline model. This shows that the improved model has a slight increase in calculation volume and inference time due to the addition of attention and x-small target detection heads, but can significantly improve detection performance. With the improvements to DSSFF, mAP50 and mAP50:95 increase by 0.9% and 0.4% on the VisDrone-2019 dataset, respectively. On the DIOR dataset, mAP50 and mAP50:95 increase by 2.1% and 2.9%, respectively, indicating that the feature fusion mechanism of DSSFF not only aids in identifying dense small objects but also refines the detection of objects at different scales, as reflected in the high-threshold mAP50:95 metric. For the enhancement of the X-small object detection head, mAP50 and mAP50:95 rise by 1.8% and 1.4% on the VisDrone-2019 dataset, and by 3.1% and 4.1% on the DIOR dataset, respectively. This demonstrates that introducing the X-small detection head boosts the model's ability to detect objects of various scales, especially small-sized ones, thus

Table 3. Comparison of detection results of mainstream methods on the DIOR dataset.

Method	*AE*	*AT*	*BD*	*BT*	*BE*	*CY*	*DM*	*EA*	*EN*	*GO*	mAP50	mAP50:95
	GR	*HR*	*OS*	*SP*	*SM*	*SK*	*TT*	*TN*	*VE*	*WL*		
ATSS	94.8	88.1	93.9	90.7	50.5	90.9	70.5	90.8	76.6	86.7	80.4	56.6
	87.7	65.3	67.1	75.0	94.4	75.1	95.2	64.8	57.6	91.6		
Deformable-DETR	90.8	85.3	92.2	85.7	55.2	92.0	70.4	90.5	82.0	85.4	78.0	50.6
	88.1	39.0	72.1	68.9	93.8	67.4	92.4	60.5	59.7	89.7		
Conditional-DETR	89.6	89.9	91.0	86.7	55.7	**94.2**	85.9	92.9	78.8	88.2	79.2	52.9
	89.8	39.8	73.9	59.8	96.3	60.6	92.3	69.5	57.6	90.4		
DDOD	94.7	91.8	94.5	91.0	55.7	90.8	77.8	94.1	81.5	88.8	82.9	59.2
	88.8	69.0	70.5	75.7	95.8	78.4	95.8	68.1	62.4	93.6		
TOOD	93.3	90.0	92.4	87.3	**62.1**	94.6	80.5	94.7	82.0	87.7	84.0	60.7
	90.0	64.4	**75.8**	81.3	94.9	83.6	94.3	72.7	67.1	92.0		
Dab-DETR	95.0	91.0	94.1	90.2	55.3	91.7	76.7	93.7	82.4	87.8	82.7	59.3
	89.6	69.1	71.0	76.1	94.6	77.7	95.4	66.9	62.2	93.3		
YOLOv6n	95.2	91.3	93.6	91.6	51.7	86.3	75.5	95.1	77.3	86.0	82.9	60.2
	88.2	73.8	67.3	92.8	96.5	82.6	96.4	71.2	56.3	89.9		
DAMO-YOLO	95.9	92.6	95.1	92.7	54.8	89.5	80.3	96.6	87.3	90.1	85.5	62.7
	90.8	75.7	70.1	94.0	95.5	85.9	97.2	72.9	61.2	92.6		
RTMDET	96.2	**94.2**	94.4	**94.3**	56.7	93.6	78.4	96.3	86.0	**90.8**	85.3	63.1
	90.8	**77.8**	71.9	77.7	96.5	79.2	96.9	**78.1**	65.0	92.0		
YOLO-MS	95.2	90.9	95.2	90.9	52.4	88.1	76.3	95.1	80.2	86.8	83.9	60.3
	89.4	74.3	68.2	93.7	96.1	85.2	96.8	70.5	60.3	92.6		
Gold-YOLO	96.2	93.2	95.8	92.9	57.2	89.9	79.8	96.8	85.9	88.5	85.6	62.7
	90.9	74.7	72.2	94.2	97.0	86.5	97.4	76.2	61.5	94.2		
ASF-YOLO	95.2	91.5	94.3	91.8	53.3	87.4	76.7	94.5	80.6	86.4	83.6	60.0
	88.4	73.6	68.6	93.2	96.3	85.4	96.4	68.0	58.2	92.2		
Baseline	94.6	89.6	94.1	90.5	51.6	87.1	75.0	93.4	77.8	83.5	82.4	58.6
	88.4	72.9	67.1	92.4	96.3	83.8	95.8	68.5	56.1	89.9		
Ours	**97.0**	92.8	**96.6**	91.7	56.6	90.0	**80.7**	**96.7**	**87.9**	89.9	**87.1**	**63.8**
	91.0	76.0	69.9	**95.3**	**97.3**	**90.6**	**97.9**	70.7	**69.7**	**95.0**		

Table 4. Ablation study on VisDrone-2019 and DIOR datasets.

Dataset	VisDrone-2019				DIOR			
Method	Pre	Rec	mAP50	mAP50:95	Pre	Rec	mAP50	mAP50:95
Baseline	43.4	31.8	31.9	18.2	87.8	75.9	82.4	58.6
DSSFF	45.4	32.8	33.3	19.3	87.0	77.3	84.5	61.5
X-small	44.1	34.1	33.7	19.6	87.9	79.1	85.5	62.7
DyHead	43.9	33.2	32.6	18.5	87.2	78.1	84.5	59.8
DSSFF+X-small	47.4	36.4	37.2	22.0	**89.0**	78.5	85.7	62.3
DSSFF+DyHead	48.7	35.3	36.6	21.6	88.3	78.7	85.0	62.2
X-small+DyHead	47.8	35.8	37.0	21.7	88.4	79.5	86.2	63.0
Ours	**52.6**	**40.0**	**42.1**	**25.2**	88.8	**80.7**	**86.6**	**63.8**

comprehensively improving both mAP50 and mAP50:95 on the two aerial image datasets. Regarding the improvements to DyHead, mAP50 and mAP50:95 on the two datasets increase by 0.7%, 0.3%, 2.1%, and 1.2%, respectively, while the FPS decreases. This indicates that DyHead, which combines size, spatial, and task attention, enhances the overall expressiveness of the model but also increases the computational load and inference time. When considering the pairwise combination of these three improvements, further accuracy gains can be achieved beyond what each single improvement provides. On the DIOR dataset, adding DSSFF and the X-small detection head surpasses the final proposed

Table 5. Experimental results of DSSFF module and different upsampling methods on VisDrone-2019 dataset.

Method	Pre	Rec	mAP50	mAP50:95	FPS
Bilinear	49.9	38.5	40.6	23.6	34.1
CARAFE [28]	50.0	38.2	40.2	23.5	35.3
Nearest	50.2	38.8	41.4	24.8	45.8
DySample (Ours)	**52.6**	**40.0**	**42.1**	**25.2**	**46.7**

Table 6. General experiments on different versions of YOLO for the DASSF method on the VisDrone-2019 and DIOR datasets.

Dataset	VisDrone-2019				DIOR			
Method	Pre	Rec	mAP50	mAP50:95	Pre	Rec	mAP50	mAP50:95
YOLOv5n [7]	40.8	31.0	30.4	17.4	86.5	74.8	81.4	57.2
YOLOv5n+DASSF	50.8	39.6	40.9	24.3	88.2	78.2	85.2	61.5
YOLOv7tiny [26]	46.4	37.3	34.4	17.6	84.5	79.3	83.7	56.4
YOLOv7tiny+DASSF	52.1	42.0	41.0	22.8	85.6	75.8	82.4	58.5
YOLOv8n [8]	45.4	32.8	33.3	19.3	87.8	75.9	82.4	58.6
YOLOv8n+DASSF	52.6	40.0	42.1	25.2	88.8	80.7	86.6	63.8
YOLOv9t [27]	43.1	31.8	32.2	18.7	81.6	72.2	78.3	53.9
YOLOv9t+DASSF	46.8	33.9	35.3	20.4	88.8	78.0	84.6	62.9
YOLOv10n [24]	42.9	33.5	33.1	18.8	88.3	77.3	84.8	61.1
YOLOv10n+DASSF	52.9	40.9	42.5	25.4	87.4	79.4	85.8	62.2

DASSF method in terms of the precision metric. However, this does not change the overall trend of improved detection accuracy.

4.4.2 Ablation experiments on different upsampling methods of DSSFF module

We conduct variant experiments on the proposed DSSFF module using different upsampling methods on the VisDrone-2019 dataset. As can be seen from the experimental results in table 5, the upsampling method using DySample is optimal in all four accuracy indicators and has the highest FPS. Compared with the original nearest sampling method, precision, recall, mAP50 and mAP50:95 increase by 2.4%, 1.2%, 0.7% and 0.4% respectively. FPS reaches 46.7. This demonstrates that the DSSFF module we implement in the model not only enhances detection performance but also decreases computational overhead.

4.5. Universal Experiment of DASSF Method

We also experiment with the proposed DASSF method on two datasets using the YOLOv5n, YOLOv7tiny, YOLOv8n, YOLOv9t, and YOLOv10n models. Table 6 presents the experimental results. Compared with YOLOv5n, YOLOv7tiny, YOLOv8n, YOLOv9t, and YOLOv10n, although the Precision of DASSF on the DIOR dataset is slightly lower than that of YOLOv10n, our proposed DASSF method surpasses the original YOLO models in terms of Precision, Recall, mAP50, and mAP50:95. This demonstrates that the proposed DASSF method can be flexibly applied to various mainstream YOLO models, showing universality while achieving better results in aerial image target detection compared to the original YOLO models.

4.6. Visualization

Fig. 5 presents a comparative visualization of two datasets using DASSF-YOLOv8 and YOLOv8n. In the figure, red circles highlight missed detections, while yellow circles indicate false detections. The analysis demonstrates that the DASSF-YOLOv8 model not only enhances accuracy in detecting densely overlapping objects, such as motorcyclists and port areas, but also excels at identifying small targets obstructed by houses and trees, as well as those in blurred images. This leads to a reduction in missed detection rates. Furthermore, the model accurately distinguishes between positive and negative samples, resulting in fewer false detections.

5. Conclusion

This study proposed an effective aerial image detection method based on dynamic-attention scale- sequence fusion, which improves the problem of low detection accuracy of small targets in aerial images and can be flexibly applied to different YOLO models. We proposed DSSFF module to reduce the amount of calculation. By incorporating additional x-small detection heads, the detection capability of small targets can be improved. Enhanced expression capabilities for various types of targets through the use of DyHead. Compared with the baseline method and other main-

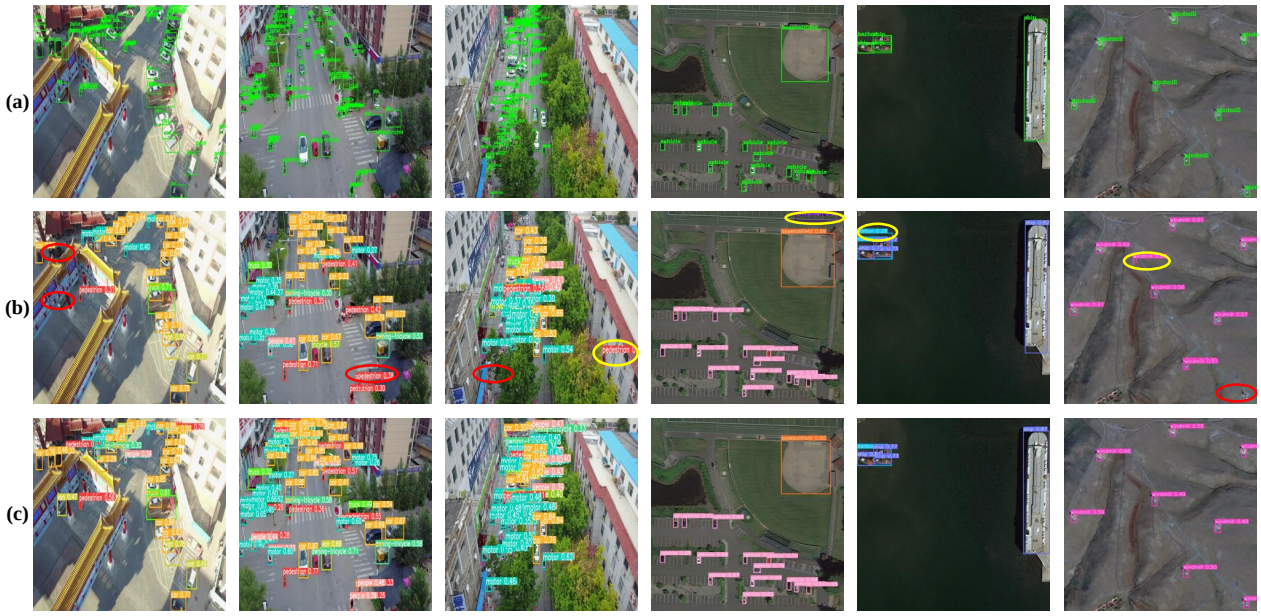


Figure 5. Comparison of detection results on VisDrone-2019 and DIOR datasets. The (a) row shows the ground truth result of the image, the (b) row shows the visualization result of the baseline model, and the (c) row shows the visualization result of our model.

stream detection methods, this approach improved detection accuracy across various challenging scenarios in aerial object detection. However, our method still has the drawback of not being fast enough during inference. In the future, we will continue to explore further lightweighting our method to ensure effective model deployment and operation even in resource-constrained environments.

Acknowledgement

This paper was supported by the National Natural Science Foundation of China (GrantNo. 42271409) and the Scientific Research Foundation of the Higher Education Institutions of Liaoning Province (GrantNo. LJKMZ20220699).

References

[1] Y. Chen, X. Yuan, R. Wu, J. Wang, Q. Hou, and M.-M. Cheng. Yolo-ms: rethinking multi-scale representation learning for real-time object detection. *arXiv preprint arXiv:2308.05480*, 2023. 6

[2] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, and F. Wu. Disentangle your dense object detector. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4939–4948, 2021. 6

[3] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382, 2021. 5

[4] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 6

[5] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021. 6

[6] Y. Jiao, S. Qiu, M. Chen, D. Han, Q. Li, and Y. Lu. Dsamgn: Graph network based on dynamic similarity adjacency matrices for vehicle re-identification. In *Pacific Rim International Conference on Artificial Intelligence*, pages 353–364. Springer, 2023. 1

[7] G. Jocher. YOLOv5 by Ultralytics, May 2020. 9

[8] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics YOLO, Jan. 2023. 6, 9

[9] M. Kang, C.-M. Ting, F. F. Ting, and R. C.-W. Phan. Asf-yolo: A novel yolo model with attentional scale sequence fusion for cell instance segmentation. *arXiv preprint arXiv:2312.06458*, 2023. 3, 6

[10] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han. Instance-aware distillation for efficient object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023. 3

[11] C. Li, L. Li, Y. Geng, H. Jiang, M. Cheng, B. Zhang, Z. Ke, X. Xu, and X. Chu. Yolov6 v3. 0: A full-scale reloading. *arXiv preprint arXiv:2301.05586*, 2023. 6

[12] H. Li, R. Zhang, Y. Pan, J. Ren, and F. Shen. Lr-fpn: Enhancing remote sensing object detection with location refined fea-

ture pyramid network. *arXiv preprint arXiv:2404.01614*, 2024. 2

[13] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 6

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[15] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 6

[16] W. Liu, H. Lu, H. Fu, and Z. Cao. Learning to upsample by learning to sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6027–6037, 2023. 4

[17] P. Luo, B. Wang, H. Wang, F. Ma, H. Ma, and L. Wang. An ultrasmall bolt defect detection method for transmission line inspection. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. 3

[18] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 6

[19] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021. 6

[20] H. Ni, H. Liu, Z. Guo, X. Wang, T. Jiang, K. Wang, and Y. Qian. Multiple visual fields cascaded convolutional neural network for breast cancer detection. In *PRICAI 2018: Trends in Artificial Intelligence: 15th Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, August 28–31, 2018, Proceedings, Part I 15*, pages 531–544. Springer, 2018. 1

[21] C. Qiao, F. Shen, X. Wang, R. Wang, F. Cao, S. Zhao, and C. Li. A novel multi-frequency coordinated module for sar ship detection. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 804–811. IEEE, 2022. 2

[22] F. Shen, X. Shu, X. Du, and J. Tang. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31th ACM International Conference on Multimedia*, 2023. 2

[23] F. Shen, Y. Xie, J. Zhu, X. Zhu, and H. Zeng. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing*, 2023. 1

[24] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024. 9

[25] C. Wang, W. He, Y. Nie, J. Guo, C. Liu, Y. Wang, and K. Han. Gold-yolo: Efficient object detector via gather-and-distribute mechanism. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[26] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. 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*, pages 7464–7475, 2023. 9

[27] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024. 9

[28] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin. Carafe: Content-aware reassembly of features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3007–3016, 2019. 9

[29] W. Weng, W. Ling, F. Lin, J. Ren, and F. Shen. A novel cross frequency-domain interaction learning for aerial oriented object detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2023. 2

[30] W. Weng, M. Wei, J. Ren, and F. Shen. Enhancing aerial object detection with selective frequency interaction network. *IEEE Transactions on Artificial Intelligence*, 1(01):1–12, 2024. 2

[31] X. Xu, Y. Jiang, W. Chen, Y. Huang, Y. Zhang, and X. Sun. Damo-yolo: A report on real-time object detection design. *arXiv preprint arXiv:2211.15444*, 2022. 6

[32] Y. Ye, Q. Huang, Y. Rong, X. Yu, W. Liang, Y. Chen, and S. Xiong. Field detection of small pests through stochastic gradient descent with genetic algorithm. *Computers and Electronics in Agriculture*, 206:107694, 2023. 3

[33] X. Ying, L. Liu, Y. Wang, R. Li, N. Chen, Z. Lin, W. Sheng, and S. Zhou. Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15528–15538, 2023. 3

[34] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 6

[35] Z. Zhao, B. Li, R. Dong, and P. Zhao. A surface defect detection method based on positive samples. In *PRICAI 2018: Trends in Artificial Intelligence: 15th Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, August 28–31, 2018, Proceedings, Part II 15*, pages 473–481. Springer, 2018. 1

[36] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 6