frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 08 December 2022
DOI 10.3389/fmars.2022.1073615

OPEN ACCESS

EDITED BY
Hong Song,
Zhejiang University, China

REVIEWED BY
Masoumeh Zareapoor,
Shanghai Jiao Tong University, China
Irfan Ahmad,
Khurasan University, Afghanistan
Zhuang Zhou,
Beijing Institute of Technology,
Zhuhai, China

*CORRESPONDENCE
Jun Qiu
qiujun@bistu.edu.cn

SPECIALTY SECTION
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

RECEIVED 18 October 2022
ACCEPTED 17 November 2022
PUBLISHED 08 December 2022

CITATION
Gao S, Liu C, Zhang H, Zhou Z and Qiu J (2022) Multiscale attention-based detection of tiny targets in aerial beach images. *Front. Mar. Sci.* 9:1073615. doi: 10.3389/fmars.2022.1073615

Multiscale attention-based detection of tiny targets in aerial beach images

Shurun Gao[1], Chang Liu[1], Haimiao Zhang[1], Zhehai Zhou[2] and Jun Qiu[1]*

[1]Institute of Applied Mathematics, Beijing Information Science and Technology University, Beijing, China, [2]Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instruments, Beijing Information Science and Technology University, Beijing, China

Tiny target detection in marine scenes is of practical importance in marine vision applications such as personnel search and rescue, navigation safety, and marine management. In the past few years, methods based on deep convolutional neural networks (CNN) have performed well for targets of common sizes. However, the accurate detection of tiny targets in marine scene images is affected by three difficulties: perspective multiscale, tiny target pixel ratios, and complex backgrounds. We proposed the feature pyramid network model based on multiscale attention to address the problem of tiny target detection in aerial beach images with large field-of-view, which forms the basis for the tiny target recognition and counting. To improve the ability of the tiny targets' feature extraction, the proposed model focuses on different scales of the images to the target regions based on the multiscale attention enhancement module. To improve the effectiveness of tiny targets' feature fusion, the pyramid structure is guided by the feature fusion module in order to give further semantic information to the low-level feature maps and prevent the tiny targets from being overwhelmed by the information at the high-level. Experimental results show that the proposed model generally outperforms existing models, improves accuracy by 8.56 percent compared to the baseline model, and achieves significant performance gains on the TinyPerson dataset. The code is publicly available *via* Github.

KEYWORDS

tiny object detection, multiscale attention, feature pyramid network, attention mechanism, unmanned aerial vehicle

1 Introduction

Target detection is the key to many computer vision applications, and its importance has gradually increased in the last decade for marine vision tasks, such as ship detection (Chen et al., 2021a; Tian et al., 2021), and maritime rescue (Varga et al., 2022), environmental monitoring (Ribeiro et al., 2019; Lieshout et al., 2020; Cheng et al., 2021). In recent years, target detection has improved tremendously as deep neural networks are trained faster and more efficiently. The combination of deep learning-based target detection technology and Unmanned Aerial Vehicle (UAV), as an image acquisition system with a large field-of-view and high efficiency, has been extensively employed for marine detection tasks with a large range and ultra-long distance. Compared with natural scene images, most of the targets in aerial marine scenes have tiny scales, and signal-to-noise ratios, and are easily swamped by background noise. Therefore, it is challenging to design a multi-scale tiny target detection model applicable to marine scenes.

The convolutional neural network-based detection model (Yu et al., 2020a; Yu et al., 2020b; Shen et al., 2019) has significantly improved the target detection task. A salient feature of the deep learning models, regarding the ability to generalize, is the quality and quantity of the dataset, and the abundant high-quality data can enhance the robustness and generalization of the model. Kisantal et al. (Kisantal et al., 2019) proposed a copy-and-paste enhancement to increase the number of samples and diversity of the tiny targets by copying and pasting images containing the tiny targets multiple times to ensure that they appear in the correct context. Chen et al. (Chen et al., 2019) proposed an adaptive resampling enhancement strategy to copy-paste the targets considering the contextual information on top of Kisantal's work, to solve the problem of context and scale mismatch in the appearance of targets, thus achieving data enhancement. This method of increasing the number of tiny targets can, to a certain extent, increase the positive samples and better optimize the model for tiny target detection. However, the gains based on the data processing have instead been constrained by the dataset.

Recently, the super-resolution reconstruction of tiny targets based on generative adversarial networks (Li et al., 2017; Na and Fox, 2018; Mehralian and Karasfi, 2018; Deng et al., 2022) has been developed. Bai et al. (Bai et al., 2018b) have proposed a multi-task generative adversarial network that feeds the super-resolution images, generated by up-sampling tiny targets, into a multi-task discriminator network that distinguishes the super-resolution images from the real images and outputs the predicted classes and bounding boxes. Noh et al. (Noh et al., 2019) have proposed a new super-resolution method at the feature level by matching the generated high-resolution features with the perceptual fields of the low-resolution features by utilizing the dilated convolution. This will help in avoiding generation of the incorrect super-resolution features owing to the perceptual field mismatch. However, the generators in the generative adversarial networks generate limited sample diversity, and hence, it is difficult to establish a balance between the generators and discriminations.

In practical applications, data enhancement methods for tiny target features may introduce new noise, which may impair the performance of the model in extracting features. Further, the super-resolution structure may complicate the end-to-end model training. To solve these problems, we designed a multiscale attention-based feature pyramid model of tiny object detection in aerial beach images. First, we addressed the problem of target information loss owing to the down-sampling in convolutional networks. The multiscale attention enhancement module (MAEM) was designed by employing self-attention to obtain the weight of the target location and retain the detailed information and the contextual information. Thus, the proposed model can improve the feature extraction of the tiny targets in the aerial large field-of-view and reduce the interference caused by the complex background. We designed a novel multiscale feature fusion module (MFFM) for the problem of inconsistent gradient computation in the Feature Pyramid Network (FPN) (Lin et al., 2017), which changes the original linear fusion, and employs the attention-guided maps to obtain the weights of the feature maps of different scales. This prevents the target from being overwhelmed by the high-level feature information while giving more semantic information to the low-level feature maps. Further, the proposed model pays more attention to the features of the tiny targets in the fusion process and improves the efficiency of the tiny target feature fusion. The proposed model has been validated on the TinyPerson dataset, and the experimental results show that the accuracy of the model, designed in this work, reaches 59.82%, which can be utilized in personnel search and rescue for seaside security. In summary, the contributions in this work are mainly in the following folds.

- We propose a novel network model for tiny target detection by introducing multiscale attention and feature pyramid networks. The detection performance of tiny targets is improved by enhancing the ability of tiny target feature extraction and fusion.
- We add an attention loss for the convolutional neural network to learn discriminative features to prevent tiny targets from being overwhelmed by complex backgrounds.
- We conducted comprehensive ablation experiments to demonstrate the impact of each of the proposed modules on detection results, and experimentally tested on the TinyPerson dataset with a significant improvement in tiny target detection accuracy over baseline.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed model framework, including MAEM, MFFM, and the loss function. The experimental results on TinyPerson datasets are reported in Section4 to validate the performance of the proposed model. Section 5 discusses the results, and Section 6 concludes this paper.

2 Related work

2.1 Small object detection

The advancement of deep learning technology has been improving the accuracy of object detection greatly, researchers search frameworks for small object detection specifically. The FPN proposed by Lin et al. (Lin et al., 2017) introduces a bottom-up, top-down network structure that achieves feature enhancement by fusing features from adjacent layers. Based on the FPN, Liang et al. (Liang et al., 2018) proposed a deep feature pyramid network using a feature pyramid structure with lateral connections to enhance the semantic features of small targets, and specialized anchors to detect small targets in high-resolution images. Nayan et al. (Nayan et al., 2020) proposed a new real-time detection algorithm for the problem that small targets tend to lose feature information after multi-layer networks. The algorithm uses upsampling and jumps connections to extract multi-scale features of different network depths during the training process, which improves the detection accuracy and speed of small target detection. Rotation equivariant feature image pyramid network (REFIPN) (Shamsolmoali et al., 2022b) improves the ability to focus on small targets in remote sensing images through scale adaptation. REFIPN uses a single detector in parallel with a lightweight image pyramid to extract features at a wide range of scales and orientations and generate regions of interest to improve the performance of small-scale object detection performance. Shamsolmoali et al. (Shamsolmoali et al., 2022a) proposed a weakly supervised approach for object detection in remote sensing images and designed a contextual fine-grained model with significant attention to different objects and target parts. Liu et al. (Liu et al., 2021a) proposed a high-resolution detection network for small targets, which improves the detection performance of small targets with reduced computational cost by using a shallow network for high-resolution images and a deep network for low-resolution images. These methods mentioned above improve the performance of small target detection to some extent.

2.2 Object detection in maritime

In comparison with target detection of natural scenes, aerial images have a wider detection range, so the obtained image field of view is often large. Lee et al. (Lee et al., 2018) modifies the 10 different ship categories in the You Only Look Once (YOLO) algorithm target classification and applies them to maritime video surveillance tasks, thus enabling real-time maritime detection. Ghahremani et al. (Ghahremani et al., 2018) proposed a CNN-based cascaded method for detecting maritime vessels, which takes the candidate target regions in the original image and performs additional processing to improve the accuracy of small target detection. Due to the high complexity of the cascaded method, it is not suitable for real-time monitoring applications. Moon et al. (Moon et al., 2020) proposed a new cascade Region-based CNN (RCNN) method to detect small targets in marine scenes. The improvement of small target detection accuracy is due to retaining its information in all layers.

Soloviev et al. (Soloviev et al., 2020) proposed two datasets for marine ship detection and evaluates the effectiveness of three target detection models, FasterRCNN (Ren et al., 2017), Region-based Fully Convolutional Network (R-FCN) (Dai et al., 2016), and Single Shot multibox Detector (SSD) (Liu et al., 2016), on this dataset. The FasterRCNN with ResNet101 as the backbone has the highest detection accuracy for large targets, but the detection accuracy for small targets is lower.

2.3 Attention-based maritime small object detection

Attention mechanism is widely used for target detection in marine scenarios due to its excellent performance. Woo et al. (Woo et al., 2018) proposed a mixed channel and spatial attention mechanism, which enhances the utilization of spatial and channel information for input features by obtaining attention weights after the spatial attention module and channel attention module are connected in series. For small target detection in the marine environment, Chen (Cheng et al., 2021) proposes a global attention module for sea-level small trash detection by adaptively fusing deep multiscale image and radar data features. Chen (Chen et al., 2021b) proposes an improved ImYolov3 based on an attention mechanism, which integrates spatial and channel attention modules into the network architecture of Yolov3, and improves the representational capability of the network by adjusting the perceptual fields in each branch network. This enables better differentiation between ships and backgrounds. Therefore, how to develop an attention mechanism in aerial image tiny target detection is a very interesting problem.

3 Materials and methods

In the UAV aerial beach images for target detection, the complex background tends to drown tiny targets, which is not conducive to the extraction of tiny target features. Further, the extraction of the contextual information in the images helps the model to differentiate between the target and the background (Zhu et al., 2021). Therefore, we designed the multiscale attention enhancement and fusion network with Swin-T (Liu et al., 2021b) as the backbone, and the network structure is shown in Figure 1. The main contributions are as follows. A key feature attention mechanism has been designed that contains the MAEM and attention loss, since the UAV aerial images contain variable and complex scenes, thus causing a large amount of redundant information in the context information extracted from the backbone. MAEM can guide the model to focus on tiny targets, thus obtaining an enhanced feature map A. Furthermore, we designed a novel feature fusion module MFFM. The height of the UAV aerial photography process varies, and the images of different scale targets have been

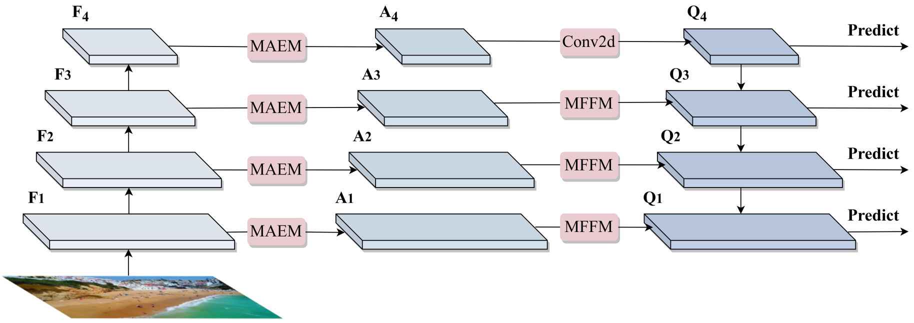

FIGURE 1
Overview of proposed AEFNet for the tiny target detection, it contains a swin-T style architecture of the feature extractor, MAEM for obtaining tiny target feature weights using a self-attention, MFFM for deep and shallow feature map fusion using attention-guided maps.

obtained. The proposed module MFFM introduces an attention mechanism in the feature fusion process to assign attention weights to the feature maps of different scales and put more attention on the tiny targets. Finally, the fused feature map *Q* has been passed through the Regions with CNN Features (RCNN) (Girshick et al., 2014) prediction network to achieve the final target location and category probability output.

3.1 Multiscale attention enhancement module

According to previous works (Yu and Koltun, 2015; Bai et al., 2018a; Zhang et al., 2021), the appropriate modeling by using contextual information has been beneficial for improving the performance of target detection. MAEM is a basic module in the network as shown in Figure 2, which preserves the detailed information of the target while obtaining contextual information. The module contains two branches, one for computing the global semantic information and the other for computing the local semantic information, which are finally computed to obtain an attention-guided map.

The local semantic information has been divided into $s{\times}s$ blocks of size $w{\times}h$ from the input original feature map F . The dependencies among the pixels in the local range have been calculated by the operation of nonlocal (Wang et al., 2018), where all blocks share weights. Thus, all the output feature maps have been gathered together to form a new local association

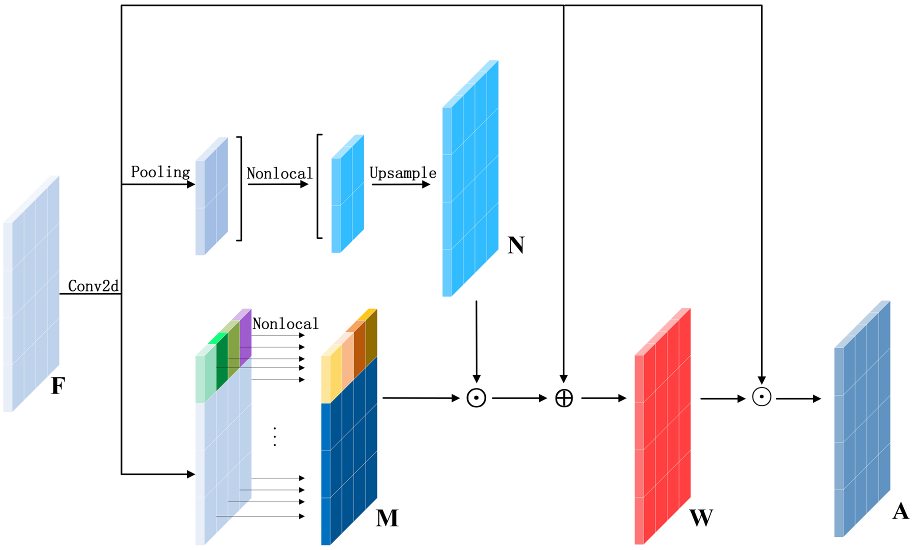

FIGURE 2
Multiscale attention enhancement module. Its upper branch and lower branch represent the global and local semantic information respectively.

feature map M . The purpose is to restrict the perceptual domain of the network to a local range, and then to use the relationships among pixels in the local range to aggregate the pixels belonging to the same class. Concomitantly, this method excludes the influence of structural noise within each patch on the target and computes the probability of the target's appearance. The design of local associations can save computational resources and speed up network training and inference.

The global semantic information has been obtained by first extracting the features of each patch from the input feature map F by adaptive pooling to obtain the pooled features with pixels $s{\times}s$, where each pixel represents a feature of each patch. Then, the contextual information among each patch has been computed by the non-local operation to travel a new guided graph N . At the global level, noise in the background and targets may have similar associations with respect to local associations, hence we use global semantic information to assist in discerning the location of targets, excluding the interference of similar targets or noise. Further, we calculate the attention-guided graph W of targets by aggregating features between the individual blocks, as shown in the following equation,

$$W_k = \alpha\sigma(N \odot M) + F_k, \tag{1}$$

where $\odot$ denotes element-wise multiplication, σ the Sigmoid activation function, and F_k the feature map at the k^{th} stage. Considering that the attention-guided map M has been employed to guide the enhanced local association features N , in this paper, the elements in M have been directly multiplied with each patch of N . Furthermore, to obtain a more effective representation, setting the learning parameter α will join the feature map F to select more effective semantic features using the adaptive nature of the network.

The attention-guided maps, generated at different stages, have different scale properties. The residual connection has been used to generate the attention-guided map into an enhanced feature map A , as shown in the following equation,

$$A_k = (1 + W_k) \odot F_k\,. \tag{2}$$

3.2 Multiscale feature fusion module

The high-level semantic and the low-level semantic have been focused on the difference in the target regions, and MFFM guides the higher layer to the shallow layer to select the appropriate features. Further, the appropriate features will be optimized to the same category, which plays the role of the feature selection. Thus, the appropriate target features in the scale range of the current layer will flow into the next layer of computation, whereas other features will be weakened and suppressed, thus enhancing the efficiency of the tiny target feature fusion. Furthermore, if the target has been detected in both the neighboring layers, the higher layer will optimize to the next layer while providing more semantic information, as follows,

$$Q_{k-1} = (I_u(Q_k) + A_{k-1}) \odot (W_{k-1} \odot I_u(W_k)), \tag{3}$$

where W_k denotes the attention-guided map of the k^{th} layer, I_u the upsampling operation, to make the adjacent layer feature maps of the same size, and Q_k the feature map after the k^{th} layer fusion.

3.3 Loss function

The method given in this paper belongs to a two-stage (Ren et al., 2017) detection model, where the first stage generates the proposal frames through a Region Proposal Network (RPN), and the second stage identifies the location of the target class through Regions of Interest (RoI). During the training process, the specific formulas of the respective loss functions of the model are shown below

$$L_{RPN} = \frac{1}{N_{cls}}\sum_{i=1} L_{cls}\left(p_{ri}, p_{ri}^{\star}\right) + \lambda_1 \frac{1}{N_{reg}}\sum_{i=1} P_{ri}^{\star} L_{reg}\left(t_{ri}, t_{ri}^{\star}\right), \tag{4}$$

$$L_{HEAD} = \frac{1}{N_{cls}}\sum_{i=1} L_{cls}\left(p_{hi}, p_{hi}^{\star}\right) + \lambda_2 \frac{1}{N_{reg}}\sum_{i=1} P_{hi}^{\star} L_{reg}\left(t_{hi}, t_{hi}^{\star}\right), \tag{5}$$

$$L_A = \frac{1}{N_{cls}}\sum_{i=1} L_{cls}\left(p_{ai}, p_{ai}^{\star}\right), \tag{6}$$

where L_{RPN} denotes the RPN loss function, L_{HEAD} the RoI Head loss function, and L_A the attention loss function. L_{cls} is the classification loss function, and the binary cross-entropy has been used to compute the classification loss, as shown in Equation (7). L_{reg} is the regression loss function, and smooth L1 has been used to compute the regression loss, as shown in Equation (8). p_{ri} denotes the prediction probability of each bounding box category, and $p_{ri}^{\star}$ denotes the truth value of each ground-truth box category. Since the role of RPN is to select the proposed box and only the foreground needs to be judged, the cross-entropy loss has been employed. Here, t_{ri} the coordinates of the bounding box, $t_{ri}^{\star}$ the coordinates of the ground-truth box, and $P_{ri}^{\star}$ determines the positive example box in the generated detection box to compute the loss. N_{cls} denotes the number of images in each small batch, and N_{reg} denotes the number of anchor box. Classification and regression losses are each normalized by N_{cls} and N_{reg} , and the parameters λ_1 and λ_2 have been used to adjust the balance of the two parts of the loss. The alternative parameter settings in L_{HEAD} and L_A are similar to those of L_{RPN} .

$$L_{cls}(p, p^{\star}) = -\log\ [pp^{\star} + (1-p)(1-p^{\star})], \tag{7}$$

$$L_{reg}(x) = \begin{cases} 0.5x^2, & |x| < 1, \\ |x| - 0.5, & |x| \geq 1. \end{cases} \tag{8}$$

Finally, the three components of the loss have been optimized by a joint loss function as

$$L = L_{RPN} + L_{HEAD} + L_A \,. \tag{9}$$

4 Experiments and results

4.1 Dataset

The numerical experiments in this paper utilize a publicly available dataset for the seaside person target detection, viz., TinyPerson (Yu et al., 2020a). As shown in Table 1.The annotation information of each sample includes the category label, bounding box, and pixel size. The size range is divided into 5 intervals: Tiny1[2, 8], Tiny2[8, 12], Tiny3[12, 20], Tiny[2, 20], and small[20, 32]. We have cropped each image with an overlap of 40 pixels to a resolution of 640 × 512 as the input to the model.

4.2 Experimental setting

We have used Swin-T as the model backbone, and the pre-training parameters loaded during the model training are those obtained by training Swin-T on the ImageNet-1K (Russakovsky et al., 2015) dataset. Experiments have been conducted based on the training and testing sample division of the dataset.

The experimental code in this paper has been implemented based on Pytorch 1.7.1, and the entire training process was deployed on AMD A40 GPU with 48 GB of memory. The code in this paper is based on the design under the MMdetection toolbox. The model is initially trained on the TinyPerson datasets with 12 epochs, and it took 48 hours in total. The gradient optimization method is AdamW (Loshchilov and Hutter, 2017), whose parameters are set as follows; i.e., weight decay at 5e-4, training batch size 4, and the learning rate initialized to 1e-4. The learning rate update is STEP, whose parameters are set as warmup iterations 1000 and warmup ratio 1e-3. The hyperparameters λ_1 and λ_2 are set as 0.6 and 1, respectively. The number of RPN proposal boxes is set as 2000 and 1000 in the training and testing phases, respectively.

To maintain consistency with the TinyPerson benchmark, the evaluation metric in this paper employs Average Precision (AP), Floating point operations (FLOPs), and Parameters (Params). Among them, AP is the evaluation index of the mainstream target detection model, the higher the value the better the model performance; FLOPs is used to measure the complexity of the model; Params is used to evaluate the number of parameters of the model.

4.3 Visualization analysis

This paper visualizes the attention heatmaps by a qualitative method to demonstrate the effect of attention loss on the model performance, which can visually show the part of the region that the model affects. According to Figure 3, the first image is the original image, the second image is the heatmap without the addition of attention loss, and the third image is the heatmap after the addition of attention loss. By comparison, the boundary around the target without adding the attention loss is blurred, and the boundary around the target after adding attention loss is clearer. This makes the model focus more on the tiny target area and avoid the interference of the environment to a certain extent, enhancing the accuracy of target detection, and thus verifying the effectiveness of the attention loss.

To further test the effectiveness of the model design, we present certain detection results in the Tiny Person dataset, including the people scenarios at sea level and on land, as shown in Figure 4.

According to the scene at sea level, the pose of people with their bodies fully exposed on surfboards varies greatly. With respect to the people swimming in the sea with only a part of their bodies exposed, the method in this paper can detect and identify the target and locate it. Furthermore, the scene of people on the land contains dense crowds, cluttered backgrounds, and different-scale crowds. Our method can still detect and identify most of the target people, thus verifying the effectiveness of our method.

4.4 Ablation study

Several sets of experiments have been set up in this paper to demonstrate the effect of patch size on the model in the MAEM. Choosing different patch sizes for different scales has different effects, as shown in Table 2.

We intend to cover both the target area and a certain background area for the selection of patch size so that the

TABLE 1 Details of TinyPerson.

Tiny Person	Training	Testing	Total
Label image	794	816	1,610
Annotations	42,197	30,454	72,651

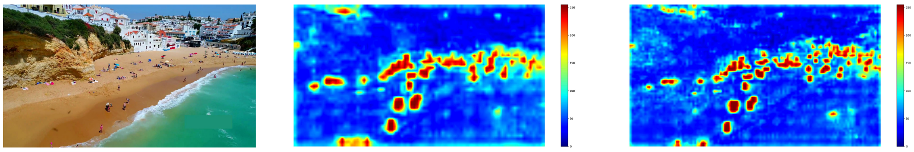

FIGURE 3
Result of attention heatmap. From left to right, the original mage, the heatmap, and the heatmap after adding attention loss.

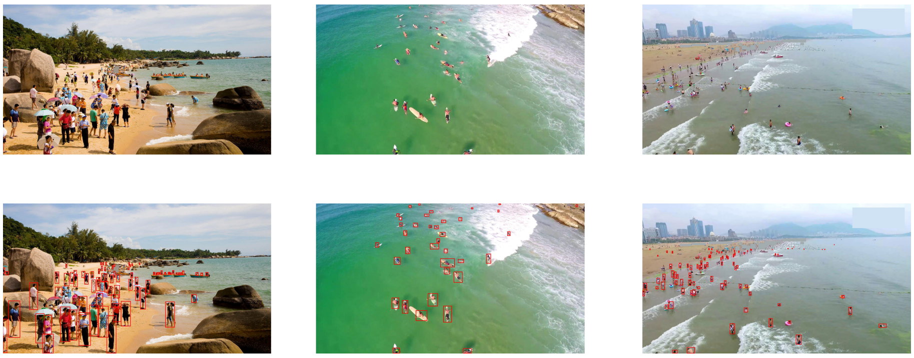

FIGURE 4
Detection results on the TinyPerson dataset. The first row is the original image of different scenes, and the second row is the detection results of the corresponding scenes.

experiments have been conducted for both the single-scale and multiscale combinations. According to Table 2, the experiments with a single size incorporate relatively little information, thus resulting in poorer results. In contrast, better results tend to have more dimensions combined in MAEM. The best results tested in the Tiny Person dataset have been for the combinations of 20,16,15, and 10.

This ablation experiment has been conducted on the Tiny Person dataset to demonstrate the effects of the MAEM, MFFM, and attention loss on the model performance. The results are listed in Table 3.

MAEM has been employed to improve the feature extraction of the tiny targets in the large field-of-view and reduce the interference generated by the complex environments, thus improving detection directly with the base feature map by 0.78% . MFFM optimizes the delivery of suitable features from the deep to shallow levels and improves the detection directly with the base feature map by 1.29% . Attention loss improves the performance by compensating for the shortcomings of the model's focus on the pixel-level classification errors, with the addition of attention loss that improves the model by 1.14% .

TABLE 2 Ablation study on patch size.

Patch Size	AP_{50}^{tiny}
5,5,5,5	57.51
10,6,5,3	57.96
15,10,8,5	59.13
20,16,15,10	59.82

4.5 Comparison to state-of-the-art methods

Table 4 shows the results of the method proposed in this paper relative to the other detection methods, including FasterRCNN

TABLE 3 Ablation study on whole network.

MAEM	MFFM	Attention loss	AP_{50}^{tiny}
×	×	×	55.54
✓	×	✓	56.32
×	✓	✓	56.83
✓	✓	×	58.68
✓	✓	✓	59.82

TABLE 4 Comparison of different methods on TinyPerson.

Method	AP_{50}^{tiny}	AP_{50}^{tiny1}	AP_{50}^{tiny2}	AP_{50}^{tiny3}	AP_{50}^{small}	AP_{25}^{tiny}	AP_{75}^{tiny}	Flops	Params
FasterRCNN	51.26	35.73	56.08	61.25	66.96	71.13	6.46	$2*10^{11}$	$51.1*10^{6}$
RetinaNet	52.43	41.38	57.68	60.54	66.85	76.32	6.73	$5.5*10^{11}$	$31.4*10^{6}$
FoveaNet	54.18	37.97	57.27	64.84	71.63	74.81	7.12	$5.4*10^{11}$	$36.0*10^{6}$
Swin-T	55.54	42.41	58.32	64.56	68.53	74.00	7.03	$2.1*10^{11}$	$44.7*10^{6}$
Yolox	57.12	44.37	59.51	66.03	70.11	76.28	8.13	$1.9*10^{11}$	$54.2*10^{6}$
PVTv2	56.38	43.15	58.73	65.39	69.82	75.12	7.82	$2.1*10^{11}$	$43.2*10^{6}$
SSPNet	58.35	46.37	60.91	66.92	71.90	77.76	8.48	$4.6*10^{11}$	$46.8*10^{6}$
OUR	**59.82**	**47.59**	**61.74**	**68.43**	**72.23**	**78.09**	**9.21**	$2.1*10^{11}$	$51.6*10^{6}$

Bold values indicates the best detection result in each column.

TABLE 5 Comparison of different methods on WSODD.

Method	mAP_{50}^{small}	mAP_{50}^{medium}	mAP_{50}^{large}	mAP_{50}^{max}
FasterRCNN	12.6	17.1	31.6	51.2
SSD	15.3	18.4	28.6	52.8
Yolov3	23.1	26.6	**41.5**	**55.9**
RetinaNet	18.3	20.1	33.8	52.8
CenterNet	10.4	24.1	29.8	43.5
Swin-T	25.8	26.4	32.7	53.6
OUR	**27.1**	**28.2**	40.3	53.7

Bold values indicates the best detection result in each column.

(Ren et al., 2017), RetinaNet (Lin et al., 2020), FoveaNet (Kong et al., 2020), Swin-T (Liu et al., 2021b), Yolox (Ge et al., 2021), PVTv2 (Wang et al., 2022) and SSPNet (Hong et al., 2022) on the TinyPerson dataset. Since some of these methods have not been previously applied to the TinyPerson datasets, all the compared methods have been tested while keeping the same configuration. The left subscript of the evaluation metric AP represents the value of the IoU in the target detection, and the right superscript represents the size of the detection target.

According to Table 4, on the TinyPerson dataset, the accuracy of our proposed method on the evaluation index AP_{50}^{tiny} is 8.56% and 7.39% higher than the anchor-based Faster-RCNN and RetinaNet methods, respectively. Further, it is 5.64% higher than the anchor-free FoveaNet method, 3.28% higher than the Transformer-based Swin-T, and 1.53% higher than the latest method SSPNet. The method in this paper shows a significant improvement compared to the baseline model with comparable Flops and Params. The proposed method in this paper achieves the best results in several other evaluation metrics. The comparison of results on the TinyPerson dataset fully illustrates the effectiveness and superiority of our method for the tiny target detection task.

4.6 The performance on water surface object detection dataset

To evaluate the detection performance of our method in marine scenarios, we conducted experiments under the Water Surface Object Detection Dataset (WSODD) (Zhou et al., 2021), as shown in Table 5. Unlike the TinyPerson dataset, the WSODD dataset contains 14 different kinds of water targets, and we selected seven algorithms (Duan et al., 2019) for testing and used mean Average Precision (mAP) as the average evaluation metric to measure the detection accuracy of all categories. In addition, the proportion of images occupied by each class of targets was divided into four parts: small (≤ 10 %), medium (10-20 %), large (20-30 %), and max (≥ 30 %). The test results show that the method proposed in this paper can achieve higher accuracy on small and medium size targets compared with other methods. The visual inspection comparison is shown in Figure 5. It shows the detection results of the baseline model and the proposed method in this paper under the same training strategy with green and red bounding boxes, respectively. According to Figure 5, it can be obtained that compared with the baseline model, our method detects significantly more ships, especially for small targets, and the recognition rate of obscured targets is significantly improved. In general, the results with different datasets show that the method proposed in this paper has good performance for small target detection in marine scenes.

5 Discussion

The main task of the tiny target detection for aerial beach images is the accurate detection and identification of targets with very few visual features of the image. However, equivalent to common scale targets, tiny targets in aerial beach images usually

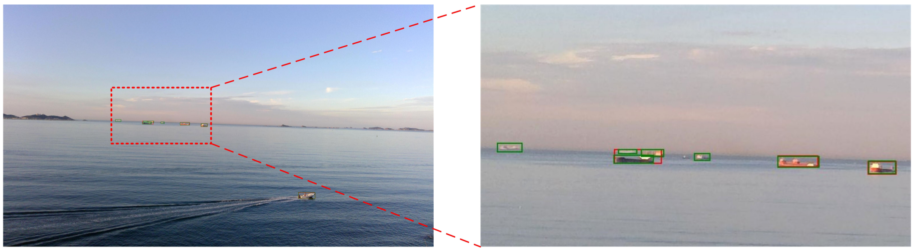

FIGURE 5
The detection result on the WSODD dataset. The red and green bounding boxes represent the detection results of the baseline and our model, respectively.

lack sufficient appearance information and are difficult to be extracted from the background as the complex scene changes (Cheng et al., 2022). Compared with the other target detection methods, the primary advantage of the proposed model in this paper is the introduction of a MAEM based on the self-attention mechanism. MAEM enables the model to effectively improve the tiny target feature extraction by learning the relationship between the target and the background. Furthermore, based on MFFM, the effectiveness of tiny target feature fusion is improved by dynamically assigning weights and using self-attention to prevent tiny targets from being overwhelmed by high-level semantic information.

Experiment results reveal that certain tiny targets cannot be accurately detected and recognized, when the scale of the target crowd in the image appears extremely high, owing to the serious tiny target occlusion situation. Therefore, exploring the optimization of the model to cope with the tiny target detection in the crowded situation will be an important direction for the subsequent work. In future work, we will also explore the theory of the attention mechanism for tiny target detection.

6 Conclusion

We proposed a multiscale attention-based feature pyramid network model, which is used for tiny target detection of aerial beach images with large field-of-view. The multiscale attention enhancement module (MAEM) in the model generates multiscale attention-guided maps to obtain contextual information while preserving the detailed information of the target. As a result, MAEM improves the ability of tiny target feature extraction in a large field-of-view and guides the model to focus more on tiny targets. The multiscale feature fusion module (MFFM) employs the attention-guided map to obtain the weights of feature maps at different scales, thus giving more semantic information to the lower-level feature maps, and effectively preventing the target from being overwhelmed by the high-level feature information. Therefore, MFFM improves the efficiency of the tiny target feature fusion. The experimental results show that the accuracy tested on the publicly available dataset Tiny Person reached 59.8% , and the ablation experiments also prove the effectiveness of each module. In future work, we will investigate the application of our proposed model to the vision processing tasks such as target traffic counting and multi-target tracking.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/ucas-vg/PointTinyBenchmark/tree/master/dataset.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s), nor the minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

SG designed the method with experiments and wrote the original draft of the manuscript supervised by JQ. HZ has overseen and led the planning and execution of research experiments. CL has been responsible for the management and coordination of the planning and execution of research activities. ZZ critically reviewed the initial manuscript and provided helpful input. All authors contributed to the article and approved the submitted version.

Funding

This work is partially supported by the National Natural Science Foundation of China (61931003, 61871042, 62171044, 12101061, 61875237), the Natural Science Foundation of Beijing (4222004), QinXin Talents Cultivation Program(Beijing Information Science and Technology University), and Beijing Great Wall Scholars Program (No.CIT&TCD20190323).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bai, Y., Zhang, Y., Ding, M., and Ghanem, B. (2018a). "Finding tiny faces in the wild with generative adversarial network," in *2018 IEEE/CVF conference on computer vision and pattern recognition* (Salt Lake City, UT, USA: IEEE), 21–30. doi: 10.1109/CVPR.2018.00010

Bai, Y., Zhang, Y., Ding, M., and Ghanem, B. (2018b). "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Computer vision – ECCV 2018*, vol. 11217 . Eds. V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (Cham: Springer International Publishing), 210–226. doi: 10.1007/978-3-030-01261-813

Cheng, Y., Xu, H., and Liu, Y. (2021). "Robust small object detection on the water surface through fusion of camera and millimeter wave radar," in *2021 IEEE/CVF international conference on computer vision (ICCV)* (Montreal, QC, Canada: IEEE). doi: 10.1109/ICCV48922.2021.01498

Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., and Han, J. (2022). Towards large-scale small object detection: Survey and benchmarks. *arXiv*. doi: 10.48550/arXiv.2207.14096

Chen, L., Shi, W., and Deng, D. (2021b). Improved YOLOv3 based on attention mechanism for fast and accurate ship detection in optical remote sensing images. *Remote Sens.* 13, 660. doi: 10.3390/rs13040660

Chen, D., Sun, S., Lei, Z., Shao, H., and Wang, Y. (2021a). Ship target detection algorithm based on improved YOLOv3 for maritime image. *J. Advanced Transportation* 2021, 1–11. doi: 10.1155/2021/9440212

Chen, C., Zhang, Y., Lv, Q., Wei, S., Wang, X., Sun, X., et al. (2019). "RRNet: A hybrid detector for object detection in drone-captured images," in *2019 IEEE/CVF international conference on computer vision workshop (ICCVW)* (Seoul, Korea (South: IEEE), 100–108. doi: 10.1109/ICCVW.2019.00018

Dai, J., Li, Y., He, K., and Sun, J. (2016). R-fcn: Object detection *via* region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* 29. doi: 10.48550/arXiv.1605.06409

Deng, C., Wang, M., Liu, L., Liu, Y., and Jiang, Y. (2022). Extended feature pyramid network for small object detection. *IEEE Trans. Multimedia* 24, 1968–1979. doi: 10.1109/TMM.2021.3074273

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). "CenterNet: Keypoint triplets for object detection," in *2019 IEEE/CVF international conference on computer vision (ICCV)* (Seoul, Korea (South: IEEE), 6568–6577. doi: 10.1109/ICCV.2019.00667

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. *ArXiv*. doi: 10.48550/arXiv.2107.08430

Ghahremani, A., Bondarev, E., and De With, P. H. (2018). "Cascaded CNN method for far object detection in outdoor surveillance," in *2018 14th international conference on signal-image technology & Internet-based systems (SITIS)* (Las Palmas de Gran Canaria, Spain: IEEE), 40–47. doi: 10.1109/SITIS.2018.00017

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE conference on computer vision and pattern recognition* (Columbus, OH, USA: IEEE), 580–587. doi: 10.1109/CVPR.2014.81

Hong, M., Li, S., Yang, Y., Zhu, F., Zhao, Q., and Lu, L. (2022). SSPNet: Scale selection pyramid network for tiny person detection from UAV images. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2021.3103069

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., and Cho, K. (2019). "Augmentation for small object detection," in *9th international conference on advances in computing and information technology (ACITY 2019)* (Sydney, Australia: Aircc Publishing Corporation), 119–133. doi: 10.5121/csit.2019.91713

Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., and Shi, J. (2020). FoveaBox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* 29, 7389–7398. doi: 10.1109/TIP.2020.3002345

Lee, S.-J., Roh, M.-I., Lee, H.-W., Ha, J.-S., and Woo, I.-G. (2018). "Image-based ship detection and classification for unmanned surface vehicle using real-time object detection neural networks," in *The 28th international ocean and polar engineering conference* (Sapporo, Japan: OnePetro), 726–730.

Liang, Z., Shao, J., Zhang, D., and Gao, L. (2018). "Small object detection using deep feature pyramid networks," in *Advances in multimedia information processing – PCM 2018*, vol. 11166 . Eds. R. Hong, W.-H. Cheng, T. Yamasaki, M. Wang and C.-W. Ngo (Cham: Springer International Publishing), 554–564. doi: 10.1007/978-3-030-00764-551

Lieshout, C., Oeveren, K., Emmerik, T., and Postma, E. (2020). Automated river plastic monitoring using deep learning and cameras. *Earth Space Sci.* 7, e2019EA000960. doi: 10.1029/2019EA000960

Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., and Yan, S. (2017). "Perceptual generative adversarial networks for small object detection," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (Honolulu, HI: IEEE), 1951–1959. doi: 10.1109/CVPR.2017.211

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (Honolulu, HI: IEEE), 936–944. doi: 10.1109/CVPR.2017.106

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. (2020). Focal loss for dense object detection. *IEEE Trans. ON Pattern Anal. AND Mach. Intell.* 42, 10. doi: 10.1109/TPAMI.2018.2858826

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "SSD: Single shot MultiBox detector," in *Computer vision – ECCV 2016*, vol. 9905 . Eds. B. Leibe, J. Matas, N. Sebe and M. Welling (Cham: Springer International Publishing), 21–37. doi: 10.1007/978-3-319-46448-02

Liu, Z., Gao, G., Sun, L., and Fang, Z. (2021a). "HRDNet: High-resolution detection network for small objects," in *2021 IEEE international conference on multimedia and expo (ICME)* (Shenzhen, China: IEEE), 1–6. doi: 10.1109/ICME51207.2021.9428241

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021b). "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF international conference on computer vision (ICCV)* (Montreal, QC, Canada: IEEE), 9992–10002. doi: 10.1109/ICCV48922.2021.00986

Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv*. doi: 10.48550/arXiv.1711.05101

Mehralian, M., and Karasfi, B. (2018). "RDCGAN: Unsupervised representation learning with regularized deep convolutional generative adversarial networks," in *2018 9th conference on artificial intelligence and robotics and 2nd Asia-pacific international symposium* (Kish Island, Iran: IEEE), 31–38. doi: 10.1109/AIAR.2018.8769811

Moon, S. W., Lee, J., Lee, J., Nam, D., and Yoo, W. (2020). "A comparative study on the maritime object detection performance of deep learning models," in *2020 international conference on information and communication technology convergence (ICTC)* (Jeju, Korea (South: IEEE), 1155–1157. doi: 10.1109/ICTC49870.2020.9289620

Na, B., and Fox, G. C. (2018). "Object detection by a super-resolution method and a convolutional neural networks," in *2018 IEEE international conference on big data (Big data)* (Seattle, WA, USA: IEEE), 2263–2269. doi: 10.1109/BigData.2018.8622135

Nayan, A.-A., Saha, J., Mozumder, A. N., Mahmud, K. R., and al Azad, A. K. (2020). Real time detection of small objects. *ArXiv*. doi: 10.35940/ijitee.E2624.039520

Noh, J., Bae, W., Lee, W., Seo, J., and Kim, G. (2019). "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *2019 IEEE/CVF international conference on computer vision (ICCV)* (Seoul, Korea (South: IEEE), 9724–9733. doi: 10.1109/ICCV.2019.00982

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Ribeiro, R., Cruz, G., Matos, J., and Bernardino, A. (2019). A data set for airborne maritime surveillance environments. *IEEE Trans. Circuits Syst. Video Technol.* 29, 2720–2732. doi: 10.1109/TCSVT.2017.2775524

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Shamsolmoali, P., Chanussot, J., Zareapoor, M., Zhou, H., and Yang, J. (2022a). Multipatch feature pyramid network for weakly supervised object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13. doi: 10.1109/TGRS.2021.3106442

Shamsolmoali, P., Zareapoor, M., Chanussot, J., Zhou, H., and Yang, J. (2022b). Rotation equivariant feature image pyramid network for object detection in optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. doi: 10.1109/TGRS.2021.3112481

Shen, W., Qin, P., and Zeng, J. (2019). "An indoor crowd detection network framework based on feature aggregation module and hybrid attention selection module," in *2019 IEEE/CVF international conference on computer vision workshop (ICCVW)* (Seoul, Korea (South: IEEE), 82–90. doi: 10.1109/ICCVW.2019.00016

Soloviev, V., Farahnakian, F., Zelioli, L., Iancu, B., Lilius, J., and Heikkonen, J. (2020). "Comparing CNN-based object detectors on two novel maritime datasets," in *2020 IEEE international conference on multimedia & expo workshops (ICMEW)* (London, UK: IEEE), 1–6. doi: 10.1109/ICMEW46912.2020.9106019

Tian, L., Cao, Y., He, B., Zhang, Y., He, C., and Li, D. (2021). Image enhancement driven by object characteristics and dense feature reuse network for ship target detection in remote sensing imagery. *Remote Sens.* 13, 1327. doi: 10.3390/rs13071327

Varga, L. A., Kiefer, B., Messmer, M., and Zell, A. (2022). "SeaDronesSee: A maritime benchmark for detecting humans in open water," in *2022 IEEE/CVF winter conference on applications of computer vision (WACV)* (Waikoloa, HI, USA: IEEE), 3686–3696. doi: 10.1109/WACV51458.2022.00374

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local neural networks," in *2018 IEEE/CVF conference on computer vision and pattern recognition* (Salt Lake City, UT, USA: IEEE), 7794–7803. doi: 10.1109/CVPR.2018.00813

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., et al. (2022). PVT v2: Improved baselines with pyramid vision transformer. *Comput. Visual Media* 8, 415–424. doi: 10.1007/s41095-022-0274-8

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "CBAM: Convolutional block attention module," in *Computer vision – ECCV 2018*, vol. 11211 . Eds. V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (Cham: Springer International Publishing), 3–19. doi: 10.1007/978-3-030-01234-21

Yu, X., Gong, Y., Jiang, N., Ye, Q., and Han, Z. (2020a). "Scale match for tiny person detection," in *2020 IEEE winter conference on applications of computer vision (WACV)* (Snowmass Village, CO, USA: IEEE), 1246–1254. doi: 10.1109/WACV45572.2020.9093394

Yu, X., Han, Z., Gong, Y., Jan, N., Zhao, J., Ye, Q., et al. (2020b). "The 1st tiny object detection challenge: Methods and results," in *Computer vision – ECCV 2020 workshops*, vol. 12539 . Eds. A. Bartoli and A. Fusiello (Cham: Springer International Publishing), 315–323. doi: 10.1007/978-3-030-68238-5_23

Yu, F., and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv*. doi: 10.48550/arXiv.1511.07122

Zhang, T., Cao, S., Pu, T., and Peng, Z. (2021). AGPCNet: Attention-guided pyramid context networks for infrared small target detection. *arXiv*. doi: 10.48550/arXiv.2111.03580

Zhou, Z., Sun, J., Yu, J., Liu, K., Duan, J., Chen, L., et al. (2021). An image-based benchmark dataset and a novel object detector for water surface object detection. *Front. Neurorobotics* 15. doi: 10.3389/fnbot.2021.723336

Zhu, X., Lyu, S., Wang, X., and Zhao, Q. (2021). "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)* (Montreal, BC, Canada: IEEE), 2778–2788. doi: 10.1109/ICCVW54120.2021.00312