IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 June 2022, accepted 8 July 2022, date of publication 12 July 2022, date of current version 25 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3190415

RESEARCH ARTICLE

MStrans: Multiscale Vision Transformer for Aerial Objects Detection

GUANLIN LU, XIAOHUI HE, QIANG WANG, FAMING SHAO, JINKANG WANG, AND LIKAI HAO

Department of Mechanical Engineering, College of Field Engineering, PLA Army Engineering University, Nanjing 210007, China

Corresponding author: Xiaohui He (gcbhxh314@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671470, and in part by the Key Research and Development Program of China under Grant 2016YFC0802900.

ABSTRACT Detecting objects in aerial images is a challenging task due to the large-scale variations and arbitrary orientations with tiny instances. A new multi-scale transformer-based aerial objects detector called MStrans is proposed in this paper to deal with the challenges in aerial detection. To detect remote instances, MStrans adopts a multi-scale patch embedding transformer (MViT) to extract the global features of the image effectively. Furthermore, to capture the different discriminant features for classification and regression branch tasks, the partial interactive fusion module (PIFM) is designed to enhance the semantic expression of the key features of classification and regression tasks by using the strategy of interactive modeling of adjacent layer features. In addition, considering that the transformer may worsen the local feature details while capturing long-distance feature dependencies, this paper designs a global to local interactive fusion module (GLIFM). It uses the advantage of convolution to extract local features to enrich the detailed information in the transformer. Experiments were carried out on DOTA and DIOR datasets, and the MStrans achieves superior detection performances compared with other approaches.

INDEX TERMS Aerial object detection, deep learning, multiscale object detection, transformer, CNN.

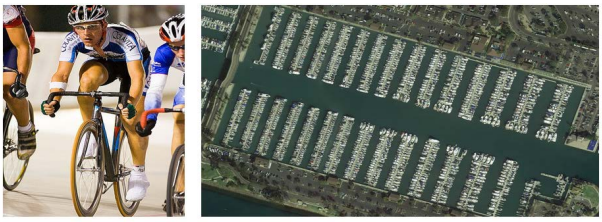
I. INTRODUCTION

In recent years, unmanned aerial vehicle (UAV) has become a piece of top-rated equipment in many applications, such as safety monitoring, aerial photography, express delivery, agricultural production, etc. UAV aerial object detection based on location and classification is widely used in crop monitoring, resource exploration, environmental monitoring, military surveillance, and other fields [1]. With the advent of the big data era, deep learning algorithms based on data support have achieved vigorous development. Many detectors based on deep learning are applied to natural image detection [2]. Although general object detection has obtained good results, the performances of these detectors in aerial image detection are far from satisfactory (both in accuracy and efficiency).

Compared with generic object detection in the MS COCO dataset [3], aerial object detection is more challenging because aerial images have an extended field of view, and many tiny instances are difficult to distinguish from the background, as shown in Fig. 1(a). To perform tiny aerial object detection better, Yang *et al.* [4] designed an end-to-end detection framework ClusDet, which combines objects clustering and detection and achieves good performance in dense small object detection. Unlike direct cutting of dense object areas by clustering, Deng *et al.* [5] designed a progressive scale-change adaptive aerial object detection framework GLSAN by adopting the global-local fusion strategy and considering the accuracy and robustness of scale-change detection. Based on the high similarity between the object distribution in aerial object detection and the object distribution in crowd counting (small scale and dense, uneven distribution), Li *et al.* [6] creatively expanded the density map from crowd counting to aerial object detection and designed DMNet, which effectively promoted the development of aerial object detection.

Similarly, Xu *et al.* [7] constructed an adaptive zoom detection network (AdaZoom) based on reinforcement learning and strategy gradient, which improved the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhongyi Guo.



(a) Instances scale in MS COCO dataset(left) and DIOR dataset(right)

(b) Instances in DIOR dataset

FIGURE 1. Challenges in aerial object detection. The length-width ratio of the instances in the red boxes changes significantly and the direction of instances in the blue boxes is randomly distributed.

multi-scale robustness of the network. To better handle the shape similarity of dense small objects in aerial images, Huang *et al.* [8] abandoned the coarse-to-fine detection strategy and adopted a fine-grained strategy to model the object distribution. Also, they designed a multi-agent detection network based on unified foreground packaging called UFPMP-Det. Although the convolution-based detection network has made great progress, convolution cannot model long-term dependencies in images. More precisely, each convolution kernel only processes a local region of the pixels in the image, and it forces the network to focus on the local mode rather than the global context, which is still not enough for accurate detection of tiny aerial objects.

Another considerable challenge of aerial object detection is the randomness of instance direction in aerial images and the significant change of instance aspect ratio, as shown in Fig. 1(b). The object detection task is usually a multi-task learning problem that includes a classification task and a regression task. The classification task is to learn the distinguishing features of the key or prominent parts of the object, while the regression task is to locate the boundary of the whole object accurately. Due to the conflict between classification and regression learning mechanisms, the spatial distribution of learning features of the two tasks may differ, resulting in a certain degree of deviation when two independent branches are used for prediction. The deviation of learning features in classification and regression branches undoubtedly increases the burden of aerial object detection based on a convolution framework. Some recent studies have overcome the above challenges by designing more powerful feature interpretation mechanisms. However, they only focus on a single class of features of aerial objects, such as rotation-invariant features [9], [10] and scale sensitive features [11], [12]. They cannot automatically extract and utilize more complex and differentiated features and cannot extract the essential features for efficient classification and regression.

Witnessing the excellent performance of transformers in learning complex dependencies between sequences, researchers are introducing transformers [13] into visual tasks. Chen *et al.* [14] designed a transformer to predict pixels and obtained competitive results with CNN for image classification. Dosovitskiy *et al.* [15] converted an image into a patch sequence, designed a vision transformer (ViT), and realized accurate recognition of multiple types of objects. In addition to primary image classification tasks, the transformer is also used to solve computer vision problems, such as object detection [16], semantic segmentation [17], and video understanding [18]. For images with a large field of view and complex scenes captured by UAVs, collecting and associating scene information from a wide area is conducive to learning the relationship between objects, improving the semantic recognition ability of objects, and alleviating category confusion.

The locality of convolution operation limits its ability to obtain global context information for a convolution network. In contrast, the transformer can globally focus on the dependencies between image feature blocks and use the attention mechanism to learn enough spatial information for efficient object detection [19]. Therefore, researchers have attempted to apply transformer to aerial object detection. Li *et al.* [20] designed a multi-level transformer network called TRD, which takes CNN as the backbone to detect objects in remote-sensing images. However, the computing resources of UAVs are often insufficient, and the two-stage detection framework increases the computing burden on the UAV platform. Zhang *et al.* [21] followed the detection strategy of a one-stage detector to design a composite network ViT-YOLO, which combined ViT and YOLO [22] and achieved good results on the VisDrone dataset [23]. In addition, considering the rich types of aerial objects and the wide range of scale changes, Zhu *et al.* [24] proposed a dense object detection network called TPH, which maintained the balance between the computing resources consumption and efficient multi-scale object detection. To sum up, because of the great potential of the transformer in visual tasks, an intuitive way to improve the detection performance is to embed the transformer into the detection framework to obtain more context information and learn more diverse feature representations.

A novel and efficient but straightforward detector with a multi-scale vision transformer called MStrans is proposed in this paper. The MStrans aims to overcome the challenges in aerial object detection (high percentage of tiny instances, multi-scale, multi-directionality, and instances of arbitrary aspect ratios). Different from the existing aerial object detector based on the transformer, MStrans abandons the traditional single-scale patch embedding in the backbone network and adopts multi-scale multi-channel patch embedding, which improves the modeling ability of long-distance multi-scale objects. Meanwhile, different from the traditional method, MStrans adopts the top-down progressive

hierarchical feature fusion strategy to cluster and fuse the key features of different branch tasks in the feature fusion stage. Moreover, to alleviate the deterioration of local detail information caused by a transformer in global feature representation, MStrans introduces a global to local feature interaction process and uses the local connectivity of convolution to stabilize the object detail features. The main contributions of this study are summarized as follows:

- A transformer-based backbone network is proposed, which makes full use of the excellent global feature expression ability of the transformer, captures rich context information, and learns more distinguishable features.
- An efficient partial interactive fusion mechanism is proposed to fuse the critical features of classification and regression branches at different levels, which guarantees the remarkable performance of MStrans.
- A global to local interactive fusion mechanism is introduced, which uses the local connectivity of convolution to preserve the local details in the global features and enrich the representation of details in aerial object detection.

II. RELATED WORK

With extensive research in recent decades, aerial object detection has developed rapidly. We focus on reviewing three research directions in this section.

A. GENERAL OBJECT DETECTION

As one of the core problems in computer vision, object detection aims to accurately locate and classify objects in images or image sequences. Traditional object detection algorithms mainly rely on manually designed features [25], but these features have weak generalization ability and poor performance in complex scenes. In recent years, with the success of convolutional neural networks (CNN) in artificial intelligence, especially in the field of computer vision, general object detection has been extensively developed. The existing detectors can be roughly divided into two categories: region-based detectors and region-free detectors. R-CNN [26], Faster R-CNN [27], and Cascade R-CNN [28] are representative region-based detectors. In comparison, SSD [29], YOLO, RetinaNet [30], and other region-free detectors do not put forward area suggestions during detection, and the efficiency is improved at the expense of accuracy. Though these general-purpose detectors perform well on natural images (e.g., images in MS COCO and images in PASCAL VOC [31]), they suffer from performance degradation when applied to wide-field, high-resolution aerial images (e.g., images in LEVIR [32] and images in DOTA [33]). Therefore, aerial object detection has recently received more and more research attention.

B. VISION TRANSFORMER

Convolution is the primary tool of traditional deep neural networks for computer vision. It has made breakthrough progress, such as solving complex image recognition tasks on high-dimensional data sets such as ImageNet [34]. However, traditional convolution has its drawbacks. For example, its receptive field is limited and cannot capture long-distance dependencies. Also, after the completion of traditional convolution training, its parameters remain fixed in a certain range, so it is not flexible and adaptive enough to dynamically adapt to any change of input. The transformer's long-distance modeling ability and dynamic self-attention learning mechanism make up for the drawbacks of convolution in visual tasks. A plethora of studies has been conducted on using transformers in visual tasks. iGPT and ViT only use transformers for image classification and explore the high-dimensional modeling method. Based on the image classification task, researchers gradually apply transformers to high-level vision tasks (object detection). Inspired by the performance of the Feature Pyramid Network (FPN) [35] in multi-scale detection, Zhang *et al.* [36] designed a transformer-based multi-scale detection network called FPT, which makes full use of cross-spatial-scale feature interactions. Unlike the method in which FPT enhances specific modules, Carion *et al.* [16] designed a complete transformer-based detection network DETR, which provides a reference for developing a complete end-to-end detector. Aiming at the high computational resource consumption of DETR, Zheng *et al.* [37] replaced the self-attention module of the pre-trained DETR model with the proposed Adaptive Clustering Transformer (ACT), which dramatically reduces the computational cost without reducing the accuracy. Similarly, Graham *et al.* [38] introduced attention bias into the patch to handle the problem of decreasing the resolution of the activation map and reduce the computational cost. With the continuous optimization of the transformer, such as reducing high computing resource consumption and enormous data dependence, more transformer-based detection networks have been proposed for object detection.

C. AERIAL OBJECT DETECTION

Despite the success of CNN in general object detection, the performance of UAV object detection can be further improved. The high percentage of tiny instances and the multi-scale multi-directional instances with arbitrary aspect ratios make this task more challenging. Inspired by the dense object detection frameworks, Chen *et al.* [39] mixed up the anchor-free detector with a re-regression module to construct a hybrid detection framework called RRNet, which has extensively promoted the development of aerial object detection on VisDrone dataset. Considering that the scale and number of minimal objects in the VisDrone data set cannot meet the training requirements of the detection framework, Wang *et al.* [40] constructed a minimal aerial object detection data set AI-TOD with an average size of 12.8 pixels, 28036 images, and 700621 instance objects. Based on this, a learning network based on multiple central points (M-CenterNet) was proposed to improve the detection performance of tiny aerial objects. The AI-TOD

data set provides data support for minimal aerial object detection. With the deepening of research, it is found that Intersection over Union (IoU) is one of the most used indicators in object detection tasks, but it is not suitable for tiny objects. Also, the detection performance of some new alternative indicators, such as Generalized Intersection over Union (GIoU), Distance-IoU (DIoU), and Complete-IoU (CIoU) are still unsatisfactory. To address this issue, Wang *et al.* [41] proposed a new indicator (Normalized Wasserstein Distance, NWD) by replacing the standard IoU with Wasserstein Distance, which improves the detection efficiency of anchor-based object detectors for tiny aerial object detection. Meanwhile, the research to handle the aerial objects of multiple scales, directions, and aspect ratios is also conducted. Yang *et al.* [42] proposed a regression loss based on Gaussian Wasserstein Distance (GWD) to successful solve boundary discontinuity and inconsistency between detection measurement and loss function in rotating object detection. GWD does not solve the boundary discontinuity and square-like problems, and it is still limited to the detection of rectangular boxes. It cannot be extended to the detection of quadrangles or polygons. Llerena *et al.* [43] proposed Gaussian bounding boxes to generate an implicit binary representation (potential rotation) ellipse. Besides, they proposed a Gaussian distribution similarity measurement method based on Hellinger distance, which improved the fit between the detector and the semantic segmentation of the objects.

These methods promote object detection on UAV images. However, due to the limited receptive field and local feature extraction, it is difficult for convolution to accurately separate tiny objects from complex backgrounds. Also, the deviation of features required for classification and regression branch tasks is less studied, which limits the performance of aerial object detection.

III. METHODS

This section presents an overview of the MStrans proposed in this paper. In general, MStrans consists of a multi-path vision transformer (MViT) backbone and two functional modules, i.e., partial interactive fusion module (PIFM) and global to local interactive fusion module (GLIFM). In Section III.A, the network structure of MStrans is introduced. Then, the details of the core components, i.e., the transformer-based backbone (Section III.B), PIFM (Section III.C), and GLIFM (Section III.D). Finally, the loss function is formulated in Section III.E.

A. NETWORK ARCHITECTURE

Similar to the visual perception range of human beings, the receptive field represents the range in which the neural network can obtain object features. The traditional convolution-based object detection network mainly expands the receptive field by convolution stacking [44]. Such an approach enables the network to obtain a larger acceptance domain at the cost of losing structural information, which limits the network's detection performance. The emergence of the transformer structure provides a new approach to new way to solve the challenges in actual aerial object detection. Fig. 2 intuitively qualitatively describes the difference between convolution and transformer in the range of receptive fields. The advantage of transformer is to capture global context information by using self-attention without losing fine-grained information, which greatly improves feature extraction ability.

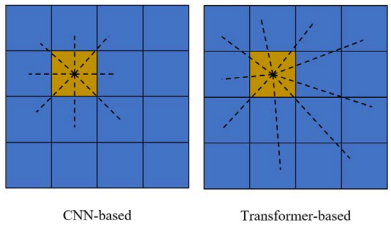


FIGURE 2. The difference between convolution and transformer in the range of receptive field.

Feature extraction of tiny aerial objects poses a challenge to the detector. Different from the previous convolution-based prediction network, this study builds a novel multi-feature progressive fusion prediction network based on the multi-path transformer, and the network structure is shown in Fig. 3. Given an input image $I \in \mathbb{R}^{W\times H\times 3}$, where W and H respectively denote the image width and height, the feature extraction branch first extracts multi-scale hierarchical features X_i with different resolutions. To better capture the critical features for classification and regression and improve the reasoning speed, this study introduces PIFM after the feature extraction branch to interactively fuse the adjacent features and form new features X_j. Finally, this study adopts GLIFM to exploit the local focusing characteristics of CNN to combine the output features of PIFM from local to global to output the prediction results.

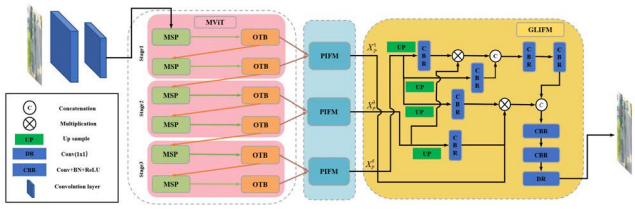


FIGURE 3. An overview of MStrans. MStrans mainly consists of the MViT backbone and two functional modules (PIFM and GLIFM), in which MSP represents the multi-scale embedding patch, and OTB denotes the object transformer block.

B. MViT BACKBONE

Based on the previous analysis of feature extraction by convolution and transformer, this study builds a feature extraction branch based on the transformer structure. Our goal is to explore an efficient backbone network for detecting multi-scale aerial objects, so this study builds a multi-level progressive architecture instead of a single-level architecture such as ViT and DETR. MStrans is proposed based on

multi-scale patch embedding and a novel dilation attention-based transformer. A three-stage design is adopted for generating different scale features. Each stage performs a multi-scale embedding patch (MSP) and consists of a series of object transformer blocks. Images need to be converted into 1D sequences before entering the transformer. As suggested by MPViT [45], the convolutional stack block has a better low-level representation without losing significant information than standard non-overlapping patch embedding. Meanwhile, a convolutional stack block with two 3×3 convolutional layers (stride 2) is adopted to generate features with a size of $\frac{H}{4} \times \frac{W}{4} \times C_0$ (C_0 is the dimension at this stage). Besides, convolution operations followed by the batch normalization and Hardswish activation function [46] with overlapping patches are adopted to exploit both fine- and coarse-grained tokens at the same feature level. Since the superimposition of continuous convolution operations under the same channel and filter size can increase the receiving field and reduce the parameter size (for example, two 3×3 convolution superpositions have the same receptive field as a 5×5 convolution, but their parameters and computational overhead are significantly reduced). As shown in Fig. 3, visual symbols can be generated with varying sizes of 3×3 and 5×5 patch and the same sequence length. Thus, given a feature $X_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, after MSP, $C_{3\times3}(X_i)$ and $C_{5\times5}(X_i)$ features can be obtained with the size of $\frac{H_i}{s} \times \frac{W_i}{s} \times C'$, where $C(\cdot)$ denotes the 2D convolution operation of size $k \times k$, stride s, and padding p; $C_{k\times k}(X_i)$ represents the output 2D token map under the MSP, and its height and width are:

$$H_i = \left\lfloor \frac{H_{i-1} - k + 2p}{s} + 1 \right\rfloor \tag{1}$$

$$W_i = \left\lfloor \frac{W_{i-1} - k + 2p}{s} + 1 \right\rfloor \tag{2}$$

The self-attention long-distance modeling mechanism is conducive to global feature extraction. However, convolution has unique advantages in dealing with low-level features. The consideration of global to local features can dramatically enhance the performance of the network in aerial object detection. Thus, this study combines MSP with the transformer embedded in dilation convolution to construct a multi-scale visual transformer called MViT to extract features. The structure of MViT is shown in Fig. 4.

Aiming at the limitation of the computing resources of the UAV platform and the actual requirement of multi-scale feature extraction, this study chooses weights-sharing dilation attention (DA) instead of standard self-attention to capture multi-scale context to achieve lightweight processing.

Specifically, multi-scale information is added to the weight of self-attention features to strengthen the feature response of attention and improve the expression ability of multi-scale features. In dilation attention, the 1×1 convolution is applied to linear projection to calculate the key and value. As for query, it is upgraded with the following operations:

$$Q = \sum\nolimits_{d\in\{1,3,5\}} SiLu\left(C\left(\bar{Q}, w_q^{k=3}, d, n\right)\right) \tag{3}$$

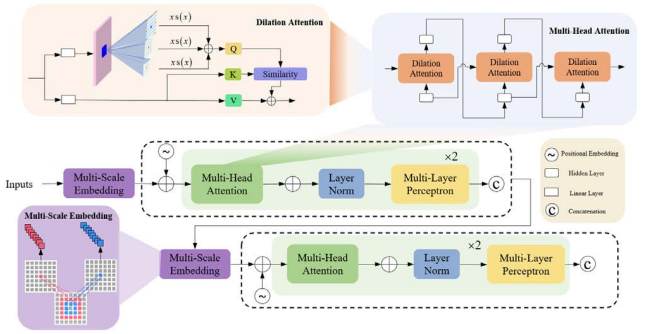


FIGURE 4. Structural diagram of the MViT backbone. In dilation attention, Q, K, and V stand for the query, key, and value in self-attention. Dilation attention calculates the multi-scale query by three depth-wise convolutions after the linear projection. These convolutions share the kernel weights but have different dilation rates, i.e., 1, 3, and 5. Their outputs are added with the weights calculated by the sigmoid function for self-calibration. This can be implemented by the SiLU function. The multi-scale information is utilized to calculate the similarity map that weights the summation of the values.

where

$$\bar{Q} = C\left(X, w_q^{k=1}, d = 1, n = 1\right) \tag{4}$$

$$SiLu(X) = X \odot sigmoid(X) \tag{5}$$

$X, Q \in \mathbb{R}^{H\times W\times C}$ are features; $w_q^k \in \mathbb{R}^{k^2\times C\times \frac{C}{n}}$ is the kernel weight; H and W denote the spatial dimensions; C is the feature dimension; k, d, and n represent the kernel size, dilation rate, and group number of the convolution, respectively. In the process of weighted summation of parallel features of different scales, this study uses the self-calibration mechanism to determine the weight of each scale through the activation intensity of the *SiLu* function. Based on this, the multi-scale information is embedded into the query between each pair of spatial positions of self-attention and the similarity calculation of the key, thus improving the ability of self-attention to extract multi-scale features.

The depth of the network structure has an important impact on the performance of the network. Meanwhile, considering the parameter usage and performance, this study uses a recursive structure in the transformer block to connect the DA in series. The operation of recursive dilation attention (RDA) is represented as follows:

$$X_{i+1} = DA(X_i + X_{i-1}) \tag{6}$$

where, i is the step, and DA is the non-linear activation function. To avoid excessive computation, this study sets the recursion depth to two.

The performance of the network depends on the characteristics of its structure. To optimize the performance of MStrans, it is necessary to analyze the internal structure of MViT. MViT is a feature extraction branch of the network, and its efficient feature extraction performance depends on the transformer block in the structure. This study investigates the recursion times of the transformer block and attempts to find the best structure configuration that balances the feature extraction branch and computing power consumption. Referring to the literature [47], three variants of MViT

TABLE 1. Detailed configurations of MViT variants.

Variant	MViT-S	MViT-M	MViT-L
Stage1	1×[MSP+OTB]	2×[MSP+OTB]	3×[MSP+OTB]
Stage2	1×[MSP+OTB]	2×[MSP+OTB]	3×[MSP+OTB]
Stage3	1×[MSP+OTB]	2×[MSP+OTB]	3×[MSP+OTB]

were designed, and the specific configuration is shown in Table 1. Meanwhile, experiments were conducted on the ImageNet dataset by employing metrics such as parameter size, FLOPs, and Top-1 accuracy, as shown in Fig. 5. It can be seen that the performance of the network increases with the network size. However, the parameter size and FLOPs present different increase trends. Specifically, MViT-S has advantages in parameter size and computational resource consumption, but the structure does not provide enough precision. Although MViT-L achieves the best accuracy among the three structures, its large parameter size and computational resource consumption are not favorable to practical detection tasks. Considering the balance of accuracy and computational resource consumption, this paper chooses MViT-M as the feature extraction branch of the network.

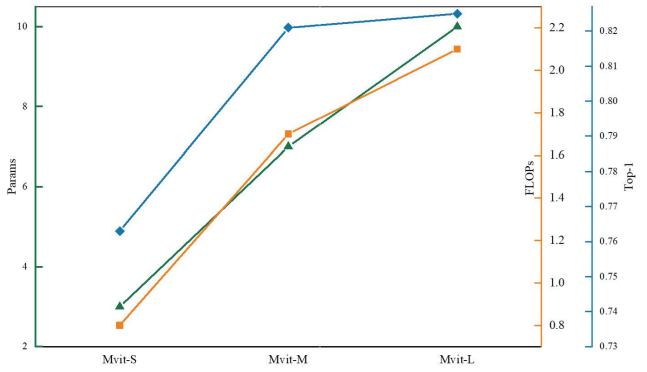


FIGURE 5. Comparison results of MViT variants.

C. PARTIAL INTERACTIVE FUSION MODULE

Regression and classification are two major tasks of object detection. Different branch tasks need different discriminant features, as shown in Fig. 6. The classification task focuses on object classification. it is necessary to weaken the influence of the position and shape of the object on the feature, i.e., to maintain scale and translation-invariance. The regression task focuses on object locating. It is necessary to ensure that the position and shape changes of the object are reflected in the features, i.e., to maintain scale and translation equality. In the regression branch, the discriminant features mainly belong to the texture information of low-level features, while in the classification branch, the discriminant features mainly belong to the abstract semantics of high-level features. Besides, compared with high-order features, low-order features have a higher spatial resolution, contribute less to performance, and require more computational resources. Previous networks usually adopted simple fusion strategies, such as feature addition or concatenation. These strategies do not consider the difference between high-level and low-level feature contributions in target detection. Although they are beneficial to the improvement of network accuracy, they cannot maximize the use of different levels of features in the regression and classification subtasks, and the influence of network accuracy improvement is not significant. From the perspective of engineering application of aerial object detection, the accuracy cannot meet the standard requirements, which will undoubtedly delay the operation process of the overall detection, prolong the operation cycle of the undertaking, and even lead to detection failure. To effectively optimize the feature fusion mechanism, and the avoid low utilization of features in classification and regression branch tasks, which limits the improvement of network accuracy, this study adopts the partial interactive fusion strategy and uses the attention mechanism to design a PIFM to interactively fuse the low-level and high-level features to maximize the use of different levels of features and improves network detection performance.

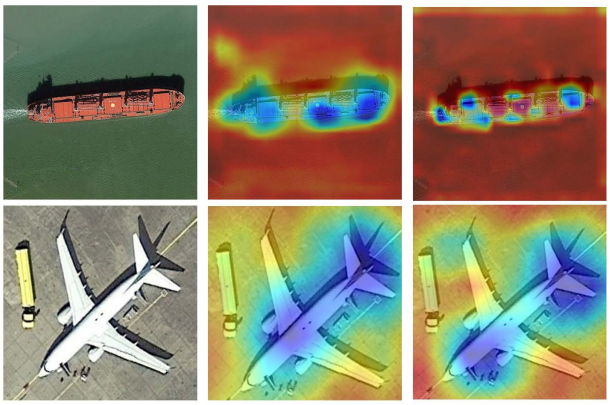


FIGURE 6. Visualizations of discriminant features for regression and classification branch tasks.

The structure of PIFM is illustrated in Fig. 7, where the lower- and higher-level features are denoted as X_l and X_h, respectively. The features are processed by two symmetric branches. The lower feature X_l is fed to a CBR component (3×3 convolution + batch normalization + ReLu activation) to generate a new feature $C(X_l)$. Based on this, the average and the maximum of the channel dimension are calculated to obtain the informative spatial attention maps A_l^{mean} and A_l^{max}.

$$A_l^{mean} = sigmoid\left(mean_d\left(C\left(X_l\right)\right)\right) \tag{7}$$

$$A_l^{max} = sigmoid\left(max_d\left(C\left(X_l\right)\right)\right) \tag{8}$$

where, $mean_d$ and max_d indicate the process to calculate the mean and max maps across the channel dimension. Similarly, A_h^{mean} and A_h^{max} are generated in the same procedure.

Then, the PIFM enhances the attention maps of the two branches as follows:

$$A_l^{mean} = P\left(A_l^{mean}, R\left(A_h^{max}\right)\right) \tag{9}$$

$$A_l^{max} = P\left(A_l^{max}, R\left(A_h^{mean}\right)\right) \tag{10}$$
$$A_h^{mean} = P\left(A_h^{mean}, R\left(A_l^{max}\right)\right) \tag{11}$$
$$A_h^{max} = P\left(A_h^{max}, R\left(A_l^{mean}\right)\right) \tag{12}$$

where $P(\cdot)$ denotes the product operation, and $R(\cdot)$ denotes the resize operation to align the dimensions of lower- and higher-level attention maps.

Next, the attention maps are concatenated and fed to the 1×1 convolution to reduce dimension. Then, considering that the classification and regression branch should focus on different features, the PIFM chooses $L1$ activation to pay more attention to the location information in the low-level features and chooses the softmax activation to pay more attention to the type attribute information in the high-level features.

$$A_l^1 = C\left(concate\left(A_l^{mean}, A_l^{max}\right)\right) \tag{13}$$
$$A_h^1 = C\left(concate\left(A_h^{mean}, A_h^{max}\right)\right) \tag{14}$$
$$\bar{X}_l = P\left(sfotmax\left(A_l^1\right), X_l\right) + X_l \tag{15}$$
$$\bar{X}_h = P\left(L1\left(A_h^1\right), X_h\right) + X_h \tag{16}$$

Finally, element-wise multiplication is performed, and the features are fed to the CBR component to obtain the final feature X_{PIFM}.

$$X_{PIFM} = C\left(mul\left(R\left(\bar{X}_l\right), C\left(\bar{X}_h\right)\right)\right) \tag{17}$$

where, $mul(\cdot)$ denotes the element-wise multiplication.

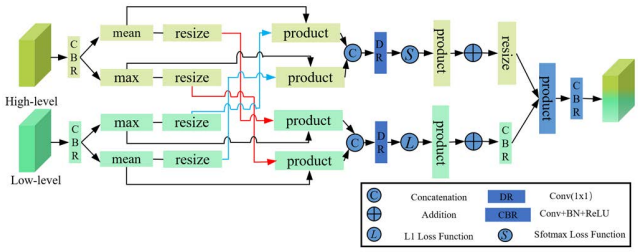


FIGURE 7. Details of the partial interactive fusion module.

D. GLOBAL TO LOCAL INTERACTIVE FUSION MODULE

Although the self-attention mechanism in a transformer can capture long-term dependencies, it is likely to ignore structural information and local relationships in each patch [48]. Meanwhile, due to some deviations in the critical features required for classification and regression branching tasks, the transformer benefits from classification deviation, thus allowing excessive focus on the essential parts of the image. By contrast, CNN benefits from regression bias and has good texture feature extraction ability. Thus, a global-to-local feature interaction module is introduced in this paper, which learns from the interaction between local and global features to deal with the preliminary fusion features $\left\{X_P^i, i = 1, 3, 5\right\}$ from the PIFM, as shown in Fig. 3.

Specifically, X_P^3 and X_P^1 are combined according to the following process:

$$h_1 = C\left(UP\left(X_p^5\right)\right) \otimes X_P^3 \tag{18}$$
$$h_2 = concate\left(h_1, C\left(UP\left(X_P^3\right)\right)\right) \tag{19}$$
$$h_3 = C\left(C\left(h_2\right)\right) \tag{20}$$

Similarly, X_P^1, X_P^3, and X_P^5 are combined according to the following process:

$$h_4 = C\left(UP\left(X_P^5\right)\right) \otimes C\left(UP\left(X_P^3\right)\right) \otimes X_P^1 \tag{21}$$

Finally, different combined features $\{h_m, m = 1, 2, 3, 4\}$ are concatenated and the final detection results Z are output.

$$Z = DR\left(C\left(C\left(concate\left(h_3, h_4\right)\right)\right)\right) \tag{22}$$

where $C(\cdot)$ denotes a sequential operation that consists of a 3×3 convolution followed by the batch normalization and ReLu function, $UP(\cdot)$ is an up-sampling operation, $\{h_m, m = 1, 2, 3, 4\}$ are combined features, and $DR(\cdot)$ is a 1×1 convolution.

E. LOSS FUNCTION

Due to small objects, dense distributions, and large-scale changes, the detection of aerial objects in large-scale scenes is challenging. Referring to [16], [47], in our framework, the Hungarian algorithm is adopted to match the prediction with the ground truth. The definition of matching cost is the same as that of loss function:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{giou}\mathcal{L}_{giou} \tag{23}$$

where, $\mathcal{L}_{cls}$ denotes focal loss; $\mathcal{L}_{L1}$ denotes L1 loss; $\mathcal{L}_{giou}$ denotes the generalized IoU loss; λ_{cls}, λ_{L1}, and λ_{giou} are the coefficients of these loss functions.

However, aerial objects usually are much smaller than normal objects. The generalized IoU loss is sensitive to the deviation of the aerial object position, resulting in similar characteristics of positive samples and negative samples and making it difficult for the network to converge. Based on this, NWD is adopted as a metric to measure the IoU loss and optimize the loss function of the network in this paper. The optimized loss function is as follows:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{NWD}\mathcal{L}_{NWD} \tag{24}$$

where, $\mathcal{L}_{cls}$ denotes focal loss; $\mathcal{L}_{L1}$ denotes L1 loss; $\mathcal{L}_{NWD}$ denotes the normalized gaussian Wasserstein distance loss; λ_{cls}, λ_{L1}, and λ_{NWD} are the coefficients of these loss functions.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, a series of comparative experiments and ablation experiments are conducted on three public and challenging aerial image data sets.

A. DATASETS

Various aerial photography detection data sets have been established according to different actual needs. The training of detection networks based on deep learning requires many high-quality data samples. As shown in Fig. 8, from the perspectives of sample number (SN), sample type distribution (STD), sample image resolution (SIR), sample labeling quality (SLQ), and instances (I) among the datasets (TAS [49], SZTALI-INRIA [50], NWPU VHR-10 [51], VEDAI [52], DIOR [53], UCAS-AOD [54], HRSC 2016 [55], and DOTA), this study finally selected DIOR and DOTA as the data sources for the experiments. Fig.10 demonstrates some samples in DIOR and DOTA, specially, DIOR contains 23463 images and 192472 instances, covering 20 object classes, including 5862 images for training, 5863 images for verification, and 11738 images for testing. DIOR alleviates some limitations of most existing data sets, such as the small number of images and object categories, insufficient image diversity, and promotes the development of aerial target detection in a data-driven mode. DOTA is one of the most significant aerial image detection benchmarks, including oriented bounding boxes and horizontal bounding boxes. DOTA contains 2806 aerial images in various scales, directions, and appearances, with a size ranging from 800 × 800 to 4000 × 4000, including 188282 instances, respectively annotated as aircraft (PL), ship (SH), large vehicle (LV), small vehicle (SV), helicopter (HC), tennis court (TC), bridge (BR), ground track field (GTF), basketball court (BC), baseball field (BD), soccer-ball field (SBF), storage tank (ST), roundabout (RA), harbor (HA), and swimming pool (SP). The whole data set is divided into a training set (1403 images), a verification set (468 images), and a test set (935 images).

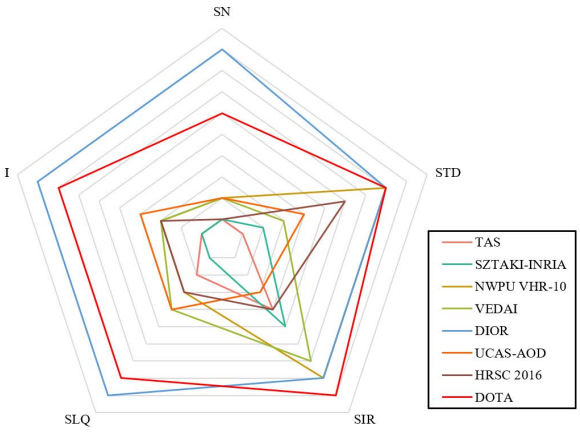


FIGURE 8. The characteristics of different data sets.

B. EVALUATIONS METRICS

Similar to the evaluation metrics used in [4, 5, 8], this study chose average precision (AP) and APs at the threshold of 0.5 (AP_{50}) and 0.75 (AP_{75}) to evaluate aerial object detection performance. Meanwhile, the variants of AP were also taken as evaluation metrics, including AP_s, AP_m, and AP_l for the object instances in a small, medium, and large size on the DOTA dataset and DIOR dataset. Furthermore, precision and recall curves were adopted to describe the detection performance from a macro perspective. The precision and recall metrics can be formulated as:

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

$$Recall = \frac{TP}{TP + FN} \tag{26}$$

where TP represents true positives; TN indicates true negatives; FN represents false negatives; FP indicates false positives.

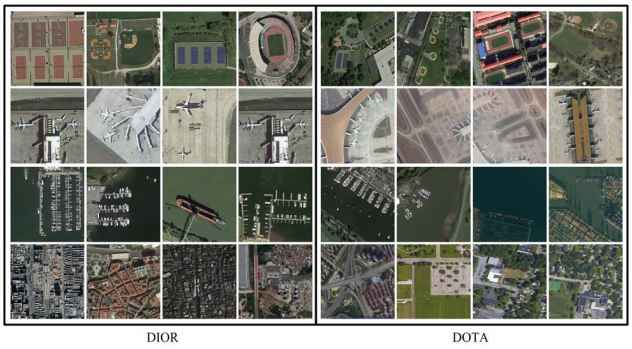


FIGURE 9. Some examples, taken from the DIOR dataset and the DOTA dataset.

C. IMPLEMENTATION DETAILS

Considering the limited computing power of the small and portable platform for aerial object detection, this study conducted experiments on a computer equipped with NVIDIA rtx3090 GPU. For the fairness of comparison, this study adopted ImageNet to pre-train all the transformer-based backbones and fine-tune them on the training sets of DOTA and DIOR. Other convolution-based detection frameworks were only trained on DOTA and DIOR training sets. Besides, random cropping and random horizontal flipping were performed to the crop the size of the DOTA dataset and DIOR dataset to 224 × 224 to reduce computing resource consumption. The AdamW optimizer was exploited to train the proposed network MStrans with an initial learning rate of 0.05, and the weight decay was set to 0.0005. In the initial training stage, 5 epochs were first chosen to warmup training, and the learning rate was updated by one-dimensional linear interpolation. After warm-up training, the cosine annealing function was employed to update the learning rate. Similarly, the coefficients of each part of the loss function were updated dynamically during training. Finally, as shown in Fig.10, the network was trained 378 epochs on the DOTA dataset with a batch size of 16 and 102 epochs on the DIOR dataset with a batch size of 32. The final learning rates were set to 0.0002 and 0.0001 on the DOTA dataset and the DIOR dataset, respectively. Moreover, $\lambda_{cls} = 1.6$, $\lambda_{L1} = 4.1$, and

TABLE 2. Comparison of different algorithms on the DOTA data set.

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
CNN-based methods						
UFPMP-Det [8]	35.7	61.2	38.8	15.2	47.2	51.6
ClusDet [4]	31.4	47.3	38.5	16.4	31.8	39.1
GLSAN [5]	32.1	54.9	31.6	17.6	35.9	41.2
Transformer-based methods						
TPH [27]	37.3	62.8	41.6	21.3	45.9	57.9
ViT-YOLO [24]	38.2	64.7	46.9	23.6	49.6	53.5
TRD [23]	39.2	65.3	48.6	25.2	51.1	52.9
LeViT [46]	41.3	63.2	47.7	24.8	50.9	53.8
MStrans (Ours)	44.5	67.5	48.4	28.3	52.7	59.8

$\lambda_{NWD} = 1.9$ on the DOTA dataset and $\lambda_{cls} = 1.5$, $\lambda_{L1} = 4.2$, and $\lambda_{NWD} = 1.7$ on the DIOR dataset.

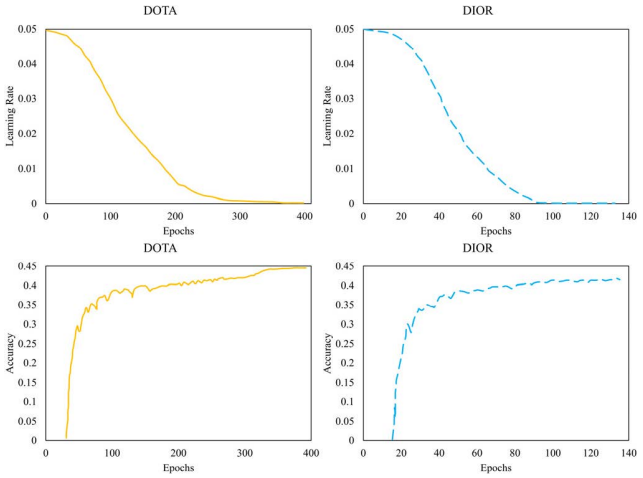


FIGURE 10. Visualization of learning rate for MStrans training on DOTA and DIOR datasets.

D. COMPARISON RESULTS

In recent years, the deep learning method has been widely used in aerial object detection. The proposed MStrans was compared with some detection frameworks, including the detection framework based on CNN (i.e., UFPMP-Det [8], ClusDet [4], and GLSAN [5]) and the detection framework based on transformer (i.e., TPH [27], ViT-YOLO [24], TRD [23], and LeViT [56]).

1) RESULTS ON THE DOTA DATASET

Our method was compared with excellent algorithms on the DOTA data set. By analyzing the results in Table 2, the transformer-based algorithms achieve better performance than CNN-based algorithms, especially for detecting small aerial objects. Also, the method proposed in this paper achieves outstanding performance among the transformer-based algorithms, and its performance in AP and AP_{50} surpasses the second place by 3.2% and 2.2%, respectively.

The transformer's long-distance modeling ability enables the transformer-based network to achieve better performance than the CNN-based network. As for MStrans, MViT effectively solves the feature extraction of dense small objects in aerial objects; PIFM partially interacts with adjacent layers' upper and lower information to process the distinguishing features required by regression subtasks and classification subtasks at different levels to improve the detection accuracy. GLIFM adopts a feature complimentary fusion mechanism to integrate global and local features, which utilizes the structural information and local information in the transformer structure and strengthens the relationship between high-level features and low-level features. Unfortunately, our method does not achieve ideal results in AP_{75}. High-precision detection is challenging for multi-scale detection tasks, and it will be the focus of our work in the future.

To compare the performances of the networks more intuitively, the detection results of different networks in five different scenarios are illustrated in Fig. 11. The first row shows the detection results of different networks in the scene of small and medium-sized objects. The results show that the transformer-based networks achieve better performance than the convolutional neural network. Unfortunately, all networks fail to detect small vehicles and shielding basketball courts in this scene. When the distribution density of the objects in the scene is reduced, the detection accuracy of the network for small objects is dramatically improved. However, false detection still occurs in different scenarios, such as the detection results of TPH-YOLOv5 in the second row, the detection results of ClusDet, GLSAN, and UFPMP-Det in the third, the detection results of LeViT in the fourth row, and the detection results of TPH-YOLOv5 and MStrans in the fifth row.

The detection of tiny and occluded objects is still challenging in object detection. Due to the unpredictable location, scale, and proportion of occlusion, the network is not robust to dense objects. Meanwhile, the network cannot extract enough object information, resulting in missed detections. The main reason for false detection is the network's lack of fine-grained classification ability, making it impossible to identify similar objects accurately. Overall, our network has performance advantages and can accurately identify aerial objects in different scenes. However, our network does not perform well in detecting similar objects and cannot accurately find subtle distinguishing features with object attributes, limiting the network's performance to a certain extent.

2) RESULTS ON THE DIOR DATASET

To further evaluate the performance, generalization ability, and stability of the proposed network, it was trained on the DIOR data set and compared with more competitive networks. Then, the detection results of the network on the DIOR dataset were classified and counted (airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf course, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill). Fig. 12 shows the P-R curves (IoU = 0.5) of different networks according to the main categories of the DIOR. The formula for calculating the accuracy rate and recall rate in Section IV.C indicates

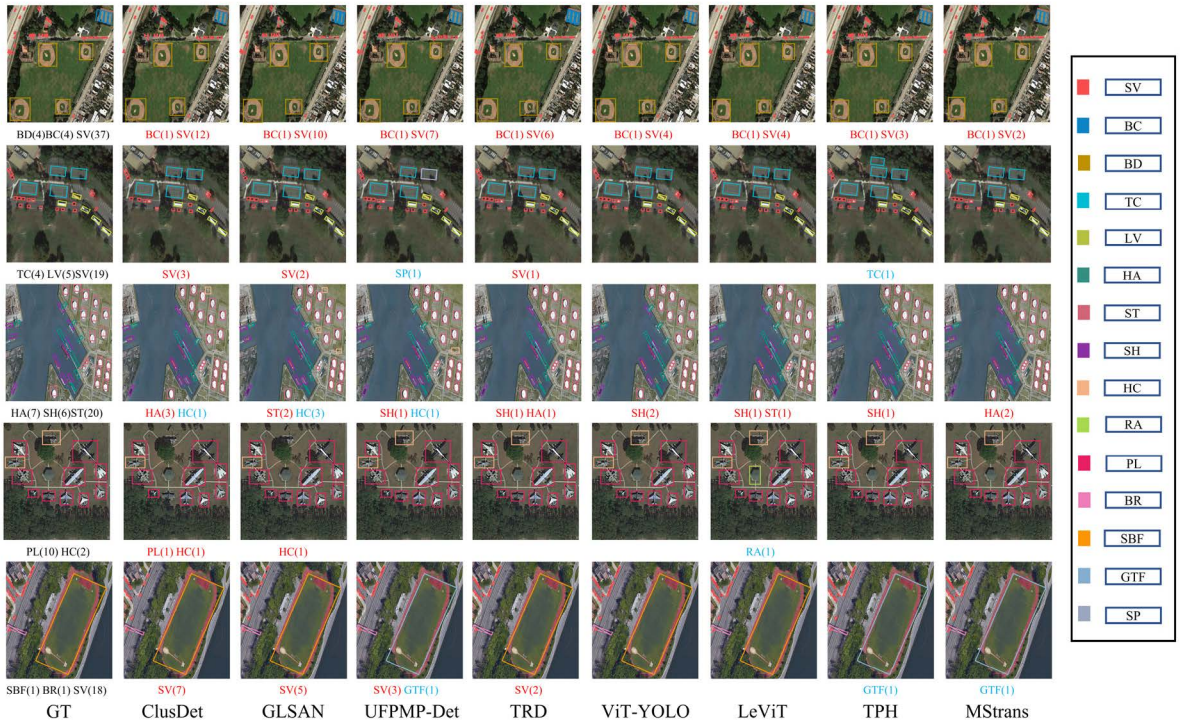


FIGURE 11. Visualizations of the detection results on the DOTA dataset. Black text indicates the number of objects in the scene; red text indicates the number of objects missing in the network; blue text indicates the number of objects missed in the network.

that the area surrounded by the P-R curve and the coordinate axis can qualitatively reflect the detection performance of the network. Overall, the proposed network's detection performance on the DIOR data set is comparable to some excellent networks. Specifically, for the accessible objects (chimney, stadium, storage tank, and wind mill) with prominent distinguishing attribute characteristics, the precision of the proposed network is not reduced by the increase of recall rate within a specific range.

For challenging objects (airplane, bridge, dam, ship, and vehicle), the network proposed in this paper still maintains the advantages of easy object detection, and its performance exceeds the second place by 3.1% to 72%. For semantic features unevenly distributed for objects with a sizeable length-width ratio, such as bridges and dams, the network performance is similar to that on the DOTA dataset, leading to false detections.

Compared with the comparison network, the network proposed in this paper effectively reduces false detections and missed detections. Specifically, for objects with variable directions, such as aircraft, MStrans pays attention to the invariance of features in rotating objects and maintains the continuous change of IoU at different rotation angles. These advantages can also be observed in the testing on the DOTA data set. MStrans adopts a transformer feature extraction structure based on multi-path, which enhances the feature semantic expression of complex objects. However, the detection accuracy of bridges (63.7%) and dams (56.8%) is still a certain distance from the requirements of actual tasks.

Similarly, MStrans has apparent advantages in detecting small and dense objects such as ships and vehicles, surpassing second place by 2.7% and 3.6%. However, these advantages are still insufficient to meet the needs of practical tasks. From the perspective of the feature interpretation mechanism, the semantic information of dense small targets often appears in the shallow feature map. With the deepening of the network, this information may disappear entirely. At this stage, the detection network is challenging to play a regular or even extraordinary performance under the condition of limited features. However, dense small targets account for a large proportion of aerial images, and their detection accuracy restricts the performance of the detection framework to a great extent. We will focus on this issue in our subsequent work.

E. ABLATION STUDY

A series of ablation experiments were conducted on the DOTA dataset to validate the effectiveness of MStrans using MViT, PIFM, and GLIFM modules for aerial objects detection. For better comparison, we adopted typical indictors: AP, AP_s, AP_m, AP_l and IT (Inference Time).

1) EFFECTIVENESS OF MViT

Many backbone networks with the expansion of the application field of object detection have been designed. The performance of the backbone network significantly affects the performance of the overall object detection framework.

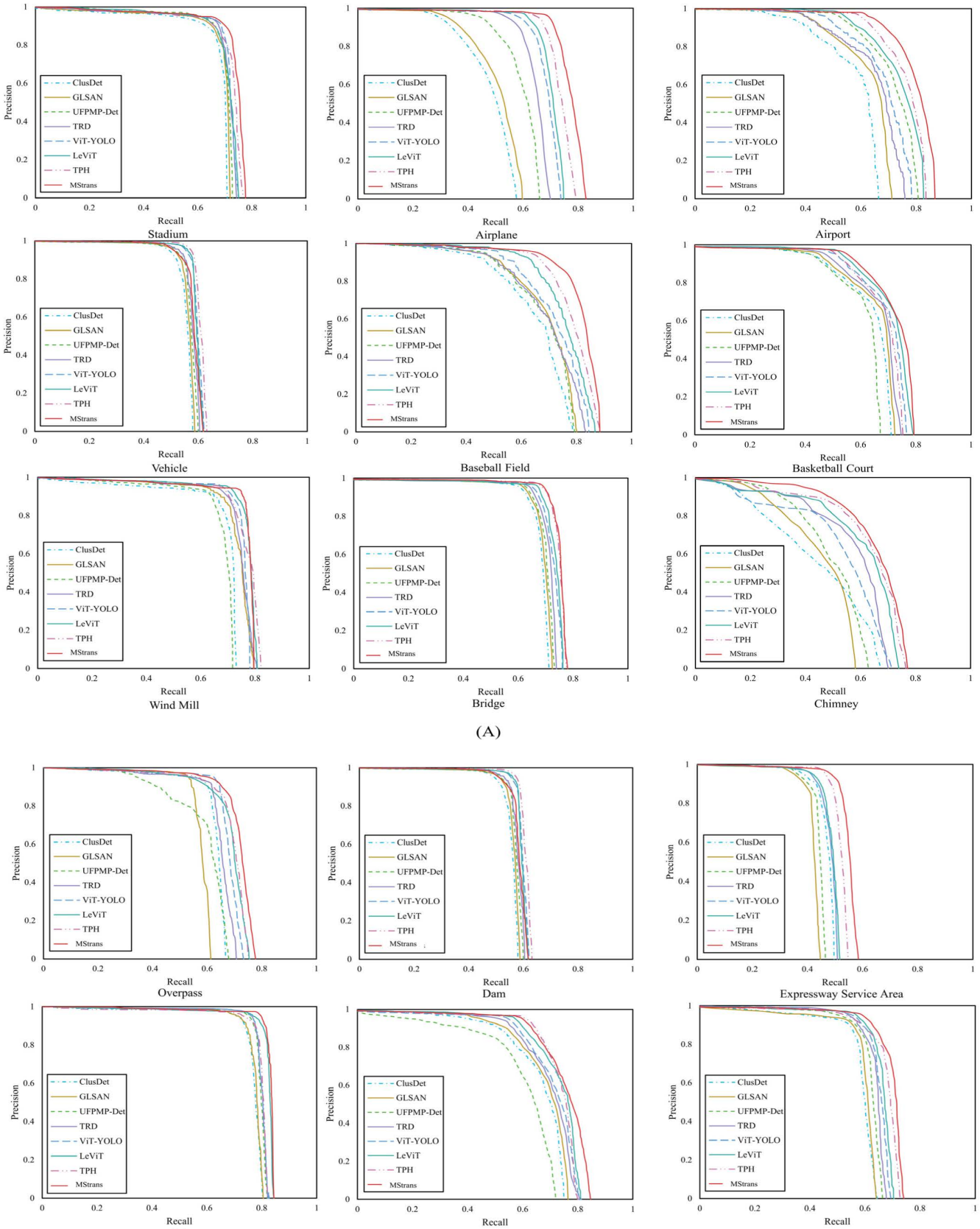


FIGURE 12. The P-R curves of different networks of the same categories on the DIOR dataset.

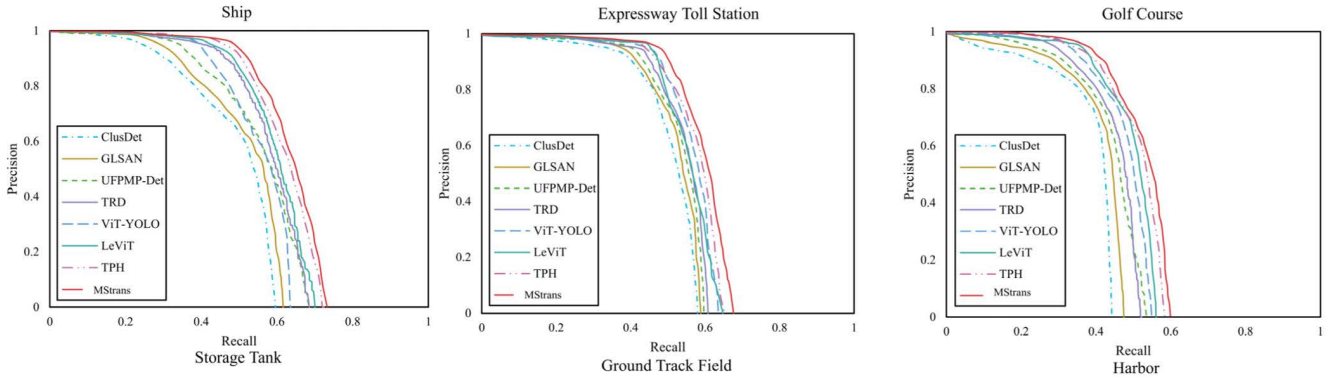


FIGURE 12. *(Continued.)* **The P-R curves of different networks of the same categories on the DIOR dataset.**

TABLE 3. Comparisons between different networks on the detection performance for the DOTA dataset.

Network	IT	AP_s	AP_m	AP_l	AP
ResNet+PIFM+GLIFM	0.196	24.3	49.7	57.7	40.5
PVT+PIFM+GLIFM	0.312	26.5	50.5	57.6	42.1
MStrans（w/o PIFM）	0.235	27.2	51.4	58.5	45.4
MStrans（w/o GIFM)	0.231	27.9	51.6	58.4	45.9
MStrans	0.277	28.3	52.7	59.7	46.8

Since PIFM and GLIFM do not have the ability of feature extraction, this study selected ResNet [57] and PVT [58], the backbone networks commonly used in object detection frameworks, as alternatives to MViT, and verified the contribution of MViT in MStrans through the performance comparison.

As shown in Table 3, MViT makes an essential contribution to the substantial improvement of the accuracy of the detection framework. Specifically, the detection accuracy of small objects increases by about 4% to 28.3% compared with the MStrans, indicating that the self-attention mechanism has advantages over convolution in the perception of small objects in aerial scenes. However, compared with ReSNet, the MViT structure is more complex, and the inference time of MStrans is increased by 0.8s than that of ResNet+PIFM+GLIFM. On the other hand, this speed is still faster than the PVT+PIFM+GLIFM, which shows that MViT has certain detection speed advantages on the premise of ensuring accuracy.

2) EFFECTIVENESS OF PIFM

PIFM is designed to handle the featured conflict between classification and regression in target detection. To verify the effectiveness of PIFM, this study replaced PIFM in the control group (w/o PIFM) with the sampling and splicing method of adjacent features. As shown in Table 3, the performances of MStrans have more significant advantages than the performances of MStrans (w/o PIFM). Especially in the small-scale object detection, the AP_s of MStrans exceed the AP_s of MStrans (w/o PIFM) by almost 1.1%, which indicates that PIFM effectively integrates the critical information in the features of adjacent layers, which ensures the overall detection accuracy of the network. Further, the critical features required by the classification and regression subtasks can effectively promote the efficient expression of object features with a small pixel ratio. Unfortunately, PIFM also increases the complexity of the network. Compared with MStrans (w/o PIFM), the inference time of MStrans increased by 0.042s, but this increase is less than that caused by MViT.

To explain the mechanism of PIFM more intuitively, some experimental results are shown in Fig. 13. Classification and regression branch tasks have different characteristics in target detection. They have feature conflicts in the same framework, which limits the detection performance of the whole framework. The two groups of experimental results were visually analyzed. In MStrans, the red box represents classification results and the blue box represents regression results. There is a small deviation between them, and they are closer to the yellow box representing the ground truth. This shows that the PIFM can undoubtedly alleviate the featured conflict between classification and regression branch tasks, which guarantees the overall detection accuracy of the framework and explains the excellent performance of MStrans on the DOTA and DIOR datasets.

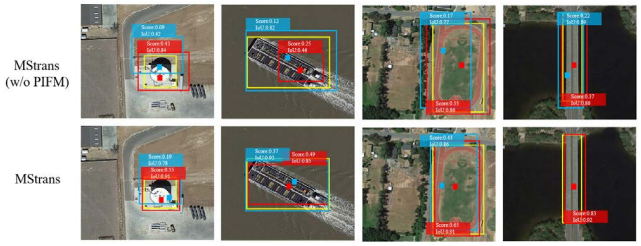


FIGURE 13. Illustration of several results. Yellow boxes indicate ground truth, and the red/blue boxes indicate the best prediction from the anchors in red patches (classification) or blue patches (regression). If the two patches coincide, only the red patches and boxes are shown.

3) EFFECTIVENESS OF GLIFM

To verify the effectiveness of GLIFM in aggregating convolution local features and global transform features, a cascade fusion operation was performed to replace GLIFM

labeled MStrans (w/o GLIFM). As shown in Table 3, MStrans performs better than MStrans (w/o GLIFM). Especially in large-scale object detection, the AP_l of MStrans is nearly 1.3% higher than the AP_l of MStrans (w/o GLIFM). The transformer pays too much attention to global features and ignores the limitations of local details. The classification features of large-scale objects are easy to obtain. However, it excessively pursues too much long-distance modeling and ignores the capture of its edge information, which is not conducive to accurately locating its region. GLIFM effectively balances the network's focus on global and local features and improves the network's ability to capture global features while learning local details. Similarly, GLIFM also increases the complexity of the network. Compared with MStrans (w/o GLIFM), the inference time of MStrans increased by 0.046s, and its inference time is consistent with that of MStrans (w/o PIFM).

To explain the mechanism of PIFM more intuitively, some experimental results are shown in Fig. 14. The hotspots in the results obtained by MStrans are more concentrated in the target and surrounding areas, and there are fewer hotspots in the background area. By contrast, MStrans (w/o GLIFM) hotspots are scattered. There are hotspots around the target and particular hotspots in the background area. Generally, the hotspots in the target detection heat map indicate that the network pays more attention to the area. Ideally, the network should eliminate the interference in the background area and focus on the object. The hotspots are also distributed on or around the target. The abnormal distribution of hotspots in the background area will reduce the detection performance of the network. Therefore, comparing the two groups of results, GLIFM complements the transformer's attention to detail information under the condition of long-distance modeling and makes the attention area more complete when the background is significantly suppressed. In this way, the hotspots are concentrated in the target and its surrounding areas, which effectively improves the recognition ability of MStrans for detail features.

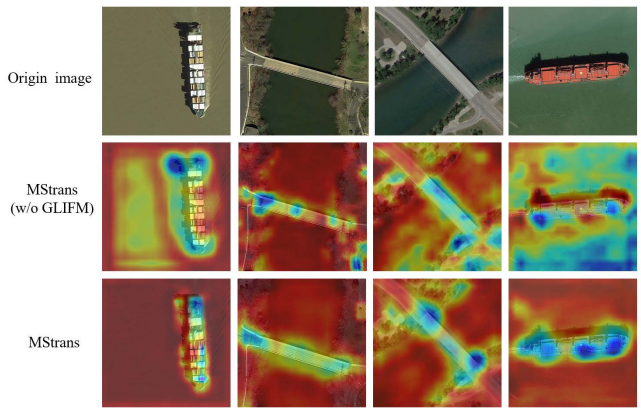


FIGURE 14. Comparison of feature maps of different networks.

To sum up, each component in MStrans contributes differently to the detection. Specifically, MViT enhances the network's ability to extract multi-scale features and promotes the overall improvement of network accuracy. PIFM mainly processes the critical features required by different branch tasks and plays an auxiliary role in improving accuracy. GLIFM mainly makes up for the network's neglect of detailed information and balances the network's attention to different scale objects. The results also conform to the mechanism of MStrans. MViT captures the original features in the image. PIFM and GLIFM mainly focus on the subsequent processing of the captured features.

On the other hand, MViT, PIFM and GLIFM all increase the complexity of the network, but MViT significantly increases the complexity of the network and reduces the inference speed. The results inspire us to pay attention to the lightweight of MViT in the follow-up research, break through the limitation of inference speed, and comprehensively improve the network performance from the two dimensions of accuracy and speed.

F. DISCUSSION

Our method mainly focuses on adopting transformer to explore a new idea to solve the two challenges (large-scale variations and arbitrary orientations with tiny instances) in aerial detection. Despite the superior performances demonstrated above, our network still has two vital limitations, which need to be deal with in futures work.

1) Fine-grained details need to enrich: The visual characteristics of targets such as ground track field and soccer ball field are similar in the aerial scene, which leads to the high similarity between classes in the aerial scene. The data-driven deep learning method has promoted the rapid development of aerial photography detection. Aerial target detection datasets represented by DOTA and DIOR are widely used in network training. However, most of these datasets are based on coarse-grained annotation information and pay insufficient attention to fine-grained details, so it is difficult for the network to learn enough detection knowledge to improve the fine-grained detection accuracy.

2) Computing needs to be controlled: The efficient feature expression ability of the deep learning method ensures detection accuracy. However, target detection is only a part of its task for small aircraft such as UAVs. It is impossible to allocate all the limited computing power of the airborne platform to target detection. The structure of the deep learning network determines the detection performance of the network. The cost of improving network performance is increased network parameters and complexity. How to optimize the network structure and control the volume of the network in future work is one of the essential factors in promoting the development of the deep learning network in the field of UAV target detection.

V. CONCLUSION

This study proposes an effective aerial object detection framework called MStrans. In the framework, MViT is used to extract feature information efficiently; PIFM can effectively alleviate the featured conflict in classification and regression

branch tasks, and GLIFM supplements the detailed information under global features and suppresses the interference in the background.

A series of ablation experiments have been conducted to verify the effectiveness of the three modules. Meanwhile, experimental results show that MStrans achieves better performance than other detection frameworks on DOTA and DIOR datasets. In particular, the AP score of MStrans on the DOTA dataset is 46.9%, which is much higher than that of other networks. The excellent performances of MStrans show that the transformer can also play a massive role in object detection outside natural language processing, which undoubtedly provides a new way for solving the challenges of aerial object detection. However, compared with the convolution-based network, the MStrans has a large number of parameters (67M) and a strong thirst for data, it still has a particular gap with the current requirements of actual detection tasks. We will focus on these issues in future research.

REFERENCES

[1] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude UAV datasets: A survey," *Image Vis. Comput.*, vol. 104, Dec. 2020, Art. no. 104046, doi: 10.1016/j.imavis.2020.104046.

[2] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Jan. 2020, doi: 10.1007/s11263-019-01247-4.

[3] 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*. Springer, 2014, pp. 740–755.

[4] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in Presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, doi: 10.1109/ICCV.2019.00840.

[5] S. Deng, S. Li, K. Xie, W. Song, X. Liao, A. Hao, and H. Qin, "A global-local self-adaptive network for drone-view object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1556–1569, 2021, doi: 10.1109/TIP.2020.3045636.

[6] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 190–191.

[7] J. Xu, Y. Li, and S. Wang, "AdaZoom: Adaptive zoom network for multi-scale object detection in large scenes," 2021, *arXiv:2106.10409*.

[8] Y. Huang, J. Chen, and D. Huang, "UFPMP-Det: Toward accurate and efficient object detection on drone imagery," 2021, *arXiv:2112.10415*.

[9] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020, doi: 10.1109/LGRS.2019.2919755.

[10] Y. Zhou, X. Liu, J. Zhao, D. Ma, R. Yao, B. Liu, and Y. Zheng, "Remote sensing scene classification based on rotation-invariant feature learning and joint decision making," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, Dec. 2019, doi: 10.1186/s13640-018-0398-z.

[11] J. Peng, M. Sun, Z.-X. Zhang, T. Tan, and J. Yan, "POD: Practical object detection with scale-sensitive network," Presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, doi: 10.1109/ICCV.2019.00970.

[12] J. Shi, H. Zhu, S. Yu, W. Wu, and H. Shi, "Scene categorization model using deep visually sensitive features," *IEEE Access*, vol. 7, pp. 45230–45239, 2019, doi: 10.1109/ACCESS.2019.2908448.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[14] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," Presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, doi: 10.1109/CVPR46437.2021.01212.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision—ECCV 2020*. Springer, 2020, pp. 213–229.

[17] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "MaX-DeepLab: End-to-end panoptic segmentation with mask transformers," Presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, doi: 10.1109/CVPR46437.2021.00542.

[18] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," Presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, doi: 10.1109/CVPR.2018.00911.

[19] Z. Sun, S. Cao, Y. Yang, and K. Kitani, "Rethinking transformer-based set prediction for object detection," Presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, doi: 10.1109/ICCV48922.2021.00359.

[20] Q. Li, Y. Chen, and Y. Zeng, "Transformer with transfer CNN for remote-sensing-image object detection," *Remote Sens.*, vol. 14, no. 4, p. 984, Feb. 2022, doi: 10.3390/rs14040984.

[21] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu, "ViT-YOLO: Transformer-BASED YOLO for object detection," Presented at the IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2021, doi: 10.1109/ICCVW54120.2021.00314.

[22] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," Presented at the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, doi: 10.1109/CVPR.2017.690.

[23] P. Zhu *et al.*, "VisDrone-DET2018: The vision meets drone object detection in image challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 1–30.

[24] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," Presented at the IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2021, doi: 10.1109/ICCVW54120.2021.00312.

[25] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, doi: 10.1109/CVPR.2014.81.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[28] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," Presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, doi: 10.1109/CVPR.2018.00644.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision—ECCV 2016*. Springer, 2016, pp. 21–37.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," Presented at the IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, doi: 10.1109/ICCV.2017.324.

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009, doi: 10.1007/s11263-009-0275-4.

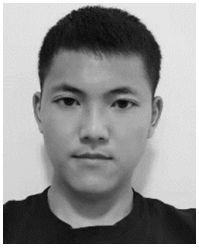
[32] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018, doi: 10.1109/TIP.2017.2773199.

[33] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," Presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, doi: 10.1109/CVPR.2018.00418.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," Presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, doi: 10.1109/CVPR.2009.5206848.

[35] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Presented at the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, doi: 10.1109/CVPR.2017.106.

[36] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Computer Vision—ECCV 2020*. Springer, 2020, pp. 323–339.
[37] M. Zheng, P. Gao, R. Zhang, K. Li, X. Wang, H. Li, and H. Dong, "End-to-end object detection with adaptive clustering transformer," 2020, *arXiv:2011.09315*.
[38] P. Stock, A. Joulin, R. Gribonval, B. Graham, and H. Jégou, "And the bit goes down: Revisiting the quantization of neural networks," 2019, *arXiv:1907.05686*.
[39] C. Chen, Y. Zhang, Q. Lv, S. Wei, X. Wang, X. Sun, and J. Dong, "RRNet: A hybrid detector for object detection in drone-captured images," Presented at the IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, doi: 10.1109/ICCVW.2019.00018.
[40] J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia, "Tiny object detection in aerial images," Presented at the 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, doi: 10.1109/ICPR48806.2021.9413340.
[41] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.
[42] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11830–11841.
[43] J. M. Llerena, L. F. Zeni, L. N. Kristen, and C. Jung, "Gaussian bounding boxes and probabilistic intersection-over-union for object detection," 2021, *arXiv:2106.06072*.
[44] Z. Zheng, Y. Zhong, A. Ma, X. Han, J. Zhao, Y. Liu, and L. Zhang, "HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 1–14, Aug. 2020, doi: 10.1016/j.isprsjprs.2020.04.019.
[45] Y. Lee, J. Kim, J. Willette, and S. Ju Hwang, "MPViT: Multi-path vision transformer for dense prediction," 2021, *arXiv:2112.11010*.
[46] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," Presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019.
[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," Presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, doi: 10.1109/ICCV48922.2021.00986.
[48] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," Presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, doi: 10.1109/ICCV48922.2021.00042.
[49] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Computer Vision—ECCV 2008* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2008, pp. 30–43.
[50] C. Benedek, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 33–50, Jan. 2012, doi: 10.1109/TPAMI.2011.94.
[51] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016, doi: 10.1109/TGRS.2016.2601622.
[52] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery : A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016, doi: 10.1016/j.jvcir.2015.11.002.
[53] 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 J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020, doi: 10.1016/j.isprsjprs.2019.11.023.
[54] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," Presented at the IEEE Int. Conf. Image Process. (ICIP), Sep. 2015, doi: 10.1109/ICIP.2015.7351502.
[55] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," Presented at the 6th Int. Conf. Pattern Recognit. Appl. Methods, 2017, doi: 10.5220/0006120603240331.
[56] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jegou, and M. Douze, "LeViT: A vision transformer in ConvNet's clothing for faster inference," Presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, doi: 10.1109/ICCV48922.2021.01204.
[57] S. Targ, D. Almeida, and K. Lyman, "ResNet in ResNet: Generalizing residual architectures," 2016, *arXiv:1603.08029*.
[58] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," Presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, doi: 10.1109/ICCV48922.2021.00061.

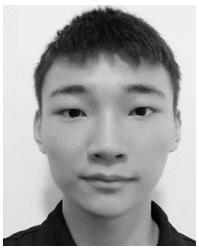


GUANLIN LU is currently pursuing the master's degree with the College of Field Engineering, PLA Army Engineering University. His research interest includes machine learning.

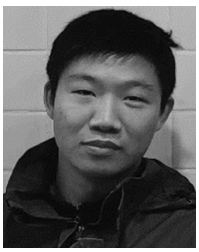
XIAOHUI HE was born in 1975. He received the Ph.D. degree from PLA Army Engineering University, China. He is a Professor with PLA Army Engineering University. His research interests include mechatronics, deep learning, and computer vision.

QIANG WANG was born in 1964. He received the Ph.D. degree from PLA Army Engineering University, China. He is a Professor with PLA Army Engineering University. His research interests include mechatronics, deep learning, and computer vision.

FAMING SHAO was born in 1978. He is an Associate Professor with PLA Army Engineering University, China. His research interests include signal processing, deep learning, and software engineering.



JINKANG WANG is currently pursuing the master's degree with the College of Field Engineering, PLA Army Engineering University. His research interest includes machine learning.



LIKAI HAO is currently pursuing the master's degree with the College of Field Engineering, PLA Army Engineering University. His research interests include numerical simulation and deep learning.