ATG-Net: Improved Feature Pyramid Network for Aerial Object Detection

Junbao Zheng, ChangHui Yang, Jiangsheng Gui*
School of Computer Science and Technology,
Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China

***Abstract*—Object detection in aerial images is gradually gaining wide attention and application. However, given the prevalence of numerous small objects in the Unmanned Aerial Vehicle (UAV) aerial images, the extraction of superior fusion features is critical for the detection of small objects. However, feature fusion in many detectors does not fully consider the specific characteristics of the detection task. To obtain suitable features for the detection task, the paper proposes an improved Feature Pyramid Network (FPN) named ATG-Net, which aims to improve the feature fusion capability. Firstly, we propose an Adaptive Tri-Layer Weighting (ATW) module that adaptively assigns weights to each layer of the feature map according to its size and content complexity. Secondly, a Triple Feature Encoding (TFE) module is implemented, which can fuse feature maps from three different scales. Finally, the paper incorporates the Global Attention Mechanism (GAM) into the network, which includes improved channel attention mechanisms and spatial attention mechanisms. The experiments are conducted on the VisDrone2020 dataset, and the result shows that the network significantly outperforms the baseline detector and a variety of popular object detectors, which significantly improves the feature fusion capability of the network and the detection accuracy of small objects.**

***Keywords*—Object detection; feature pyramid network; adaptive tri-layer weighting; triple feature encoding; global attention mechanism**

I. Introduction

Object detection technology in the UAV capture scene is rapidly advancing, and it plays an important role in the fields of power line inspection, crop analysis, military security [1], [2], [3], and so on. With the development of deep learning, especially convolutional neural networks [4], [5], [6], [7], [8], [9], the performance of object detection has been greatly improved. The detectors contain three main components: backbone, neck, and head. The primary function of the backbone is feature extraction. The mainstream architectures include VGG [10], ResNet [11], DenseNet [12], MobileNet [13], EfficientNet [14], CSPDar-knet53 [15], and SwinTransformer [16], which have been relatively mature. The main function of the neck network is multi-scale feature fusion, feature enhancement, and integration of contextual information. It plays a crucial role in the object detection task. The role of the detection head is to parse the fused feature output from the neck network, including object localization and bounding box regression. The performance of the detection head is largely dependent on the quality of the fused features. Therefore, the design of an effective necking network has a decisive impact on improving the performance of the entire detection system.

*Corresponding author

A widely adopted neck network is to build a feature pyramid network (FPN) [17], which consists of top-down paths and adds lateral connections to the network to achieve the fusion of multi-scale features, enabling the model to better understand and capture the semantic information of the object at different scales. The FPN network usually up-samples the high-level semantic feature maps and combines them with low-level features through simple summation. However, this approach does not adequately address the semantic gaps and dissimilarities between the features, thereby limiting the network's ability to generate highly discriminative features. Furthermore, fusing only high-level and low-level features cannot fully leverage the contextual information of small objects. Such a structure is limited in its ability to capture the fine details of small objects, leading to inaccurate inferences of their locations and categories, which ultimately diminishes overall object detection accuracy.

In recent years, various FPN networks have been proposed to enhance the multi-scale feature fusion capabilities. PANet [18] augmented FPN with a bottom-up path enhancement, allowing information from lower layers to be directly transferred to higher layers, thereby enhancing the flow of information. Bi-FPN [19] proposed a bidirectional cross-scale connectivity structure. This structure enhances feature fusion by adding top-down paths to the FPN. Additionally, it introduces more lateral connections between different levels. These connections improve the fusion of features. Zhang Y et al. [20] proposed a feature pyramid network that combines top-down and bottom-up approaches. By integrating these two architectures, feature maps with richer semantic information and conducive to object detection can be obtained. EFPN [21] designed a feature texture transfer module, which endows the extended feature pyramid with reliable details, extending the original FPN to specialize in small object detection for high-resolution images. SAFPN [22] designed an efficient feature pyramid network for crowded human detection, integrating a refined HS-block into the original FPN to mitigate the effects of scale variations introduced by crowds. With this structure, a single level of features can encompass more receptive fields, accommodating objects at different scales. Although the methods can obtain rich semantic information, they perform a simple summation when fusing low-level and high-level semantics without considering the varying degrees of importance among the features. As a result, they fail to generate highly discriminative features. Additionally, fusing only high-level and low-level semantic features does not fully utilize the contextual information.

Considering cross-layer feature fusion, new design schemes have been proposed. CFPN [23] is a novel cross-layer feature

pyramid network that aggregates multi-scale feature maps and then assigns the aggregated features to the corresponding layers. This enables direct cross-layer communication, improving the asymptotic fusion in salient object detection and yielding better feature maps. However, the method only scales the weights of different feature layers with scaled weights and does not further fuse the feature layers to generate highly discriminative features. ImFPN [24] proposed an improved feature pyramid network based on a similarity fusion module and an attention module, which can fuse different features to accommodate instances of varying sizes. However, the design of the fusion module neglects the differences in the relative importance of the feature maps and, to some extent, increases the computational burden.

In order to solve the above problems, this paper specifically designs a feature pyramid network named ATG-Net for aerial image detection. Firstly, in order to better utilize the contextual information of multi-scale features, a Triple Feature Encoding (TFE) module is proposed to fuse large, medium, and small scale feature maps. Considering that feature maps of different sizes may have different importance in object detection, this paper proposes an Adaptive Tri-Layer Weighting (ATW) module that is able to adaptively predict a set of weights for feature maps of different sizes. Considering that the attention mechanism can make the network more focused on the features of small objects, the Global Attention Mechanism (GAM) [20] is integrated into the network. In the following, Section II outlines the relevant research and studies. Section III details the methodologies of ATW, TFE, and GAM. The detailed comparative experiments and visual analysis are provided in Section IV. Section V concludes with a discussion of the advantages and limitations of the proposed model.

II. Related Work

A. Object Detectors

Contemporary object detectors can be roughly divided into two categories according to the detection process: one-stage and two-stage detectors. One-stage detectors directly predict the class and location of objects within an image. While these detectors offer higher computational efficiency, their accuracy is generally lower compared to alternative approaches. RetinaNet [25] overcomes the obstacle of sample imbalance by introducing focus loss and improves the detection precision. SCA-YOLO [26] proposes a multilayer feature fusion algorithm. In this approach, the single-stage object detection algorithm YOLOv5 is embedded with two newly proposed models and utilizes an adaptive feature fusion network. This enhances the network's feature representation capabilities, significantly improving the detection accuracy of small objects. ASF-YOLO [27] proposes a framework based on attentional scale sequence fusion, which combines both spatial and scale features for accurate and fast cellular instance segmentation. Compared with one-stage detectors, two-stage detectors pursue better detection accuracy at the expense of speed. The R-CNN family of detectors [28], [29] employs a Region Proposal Network (RPN) to generate high-quality candidate anchors, which are then classified and localized. This design enhances the precision of object detection. Double-head R-CNN [30] respectively uses fully connected head and convolutional head for classification and bounding box regression, achieving excellent detection performance.

B. Attention Mechanism

The application of the attention mechanism in object detection has been proven to be extremely effective, which enables the model to focus on the most important areas in the image, thereby improving the accuracy and efficiency of object detection. Squeeze-and-Excitation Networks (SENet) [31] automatically calibrates the responses of feature channels by explicitly modeling the dependencies between the feature channels. By recalibrating the responses of the channels, the model can more effectively leverage the available features. Convolutional block attention module (CBAM) [32] is a straightforward yet effective attention mechanism for feed-forward convolutional neural networks. It generates attention maps independently along the channel and spatial dimensions, thereby enabling adaptive feature refinement. Inspired by CBAM, the Global Attention Mechanism (GAM) [33] enhances the performance of deep neural networks by mitigating information loss and strengthening global interaction representation. Additionally, it incorporates 3D alignment using a multilayer perceptron for channel attention and integrates a convolutional spatial attention submodule. The global attention mechanism is able to amplify the cross-dimensional interactions and capture important features in all three dimensions (channel, spatial width, and spatial height), better preserving the effective information of the original features.

III. Approach

The proposed ATG-Net network (see Fig. 1) consists of a Tri-Layer Weighting Module (ATW), a Triple Feature Encoder Module (TFE), and a Global Attention Mechanism (GAM). ATW enhances the fusion of small, medium, and large-scale features by improving their mechanical properties. It is capable of adaptively predicting a set of weights based on the significance of each feature level for effective aggregation. TFE effectively captures localized fine features of small objects, enabling the integration of local and global information to produce fused features that are better suited for small object recognition. GAM improves the performance of deep neural networks for detecting small objects by reducing information reduction and amplifying the global interaction representation.

A. Adaptive Tri-Layer Weighting Module

Directly fusing feature maps may lead to the loss of important information. Different feature maps may contain distinct types of information, and directly adding or concatenating them can result in certain key features being masked or weakened. Different feature maps may capture distinct features, some of which may be contradictory. Directly fusing these feature maps can lead to feature conflicts, making it difficult for the model to learn effective representations. In order to solve the above problems, the paper proposes an adaptive three-layer weighting (ATW) module that can adaptively predict a set of weights based on the importance of the features for each level. This enables the generation of salient features that are more favorable for small object detection. Fig. 2 illustrates the ATW model structure. Here, C denotes the number of channels, R denotes the feature resolution,

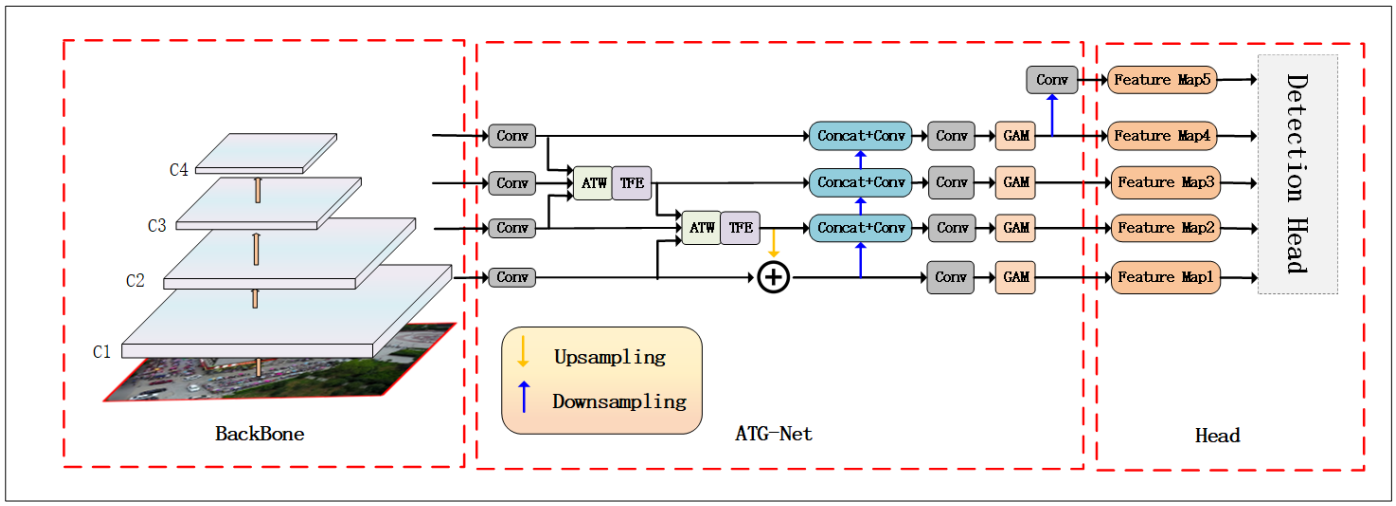

Fig. 1. The framework of ATG-Net. It consists of a backbone network, ATG-Net, and a detection head.

and FC denotes the fully connected layer. Large, Medium, and Small refer to the large-size, medium-size, and small-size feature maps, respectively. First, the features of large, medium, and small sizes are convolved by 1x1 to adjust the number of channels. Secondly, ATW employs global average pooling on each feature map to compress spatial information, resulting in a numerical value for each channel. The channel information is then concat. Finally, the concatenated features are passed through two fully connected layers to generate weight information for the three features. Formally, each layer is characterized by $X_n \in C_n \times R^{H_n \times W_n}$, and ATW computes the channel-wise global representation of $Z \in R^{C \times 1}$ by the following formula:

$$Z = \|_{n=1}^{N} z_n = \|_{n=1}^{N} \left\{ \frac{1}{H_n \times W_n} \sum_{i=1}^{H_n} \sum_{j=1}^{W_n} X_n(i,j) \right\}. \quad (1)$$

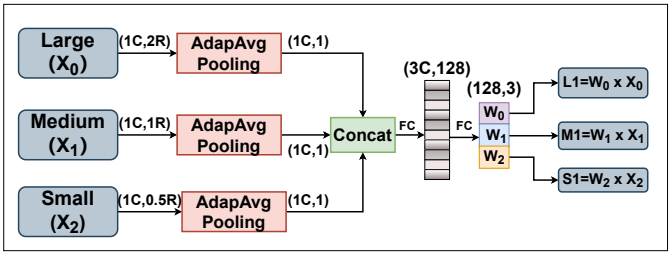

Fig. 2. The structure of the ATW module.

Where $\|$ represents the concatenation function, $C = \sum_{n=1}^{N} C_n$ represents the number of channels represented globally, n ∈ 0,1,2. N represents the number of the feature map, and $X_n(i,j)$ represents represents the feature value at the position (i,j) of the n-th feature map. This paper attempts to make use of aggregate information Z to focus on the features of each level on the significant region rather than the overall feature map. The integrated information Z is passed through two linear transformations to obtain the assigned weight $W \in R^{N \times 1}$.

$$W = FC_2(Relu(FC_1(Z))) \quad (2)$$

As shown in Fig. 2, the symbol W_n represents the nth element of W, and $\times$ denotes the scalar multiplication between X_n and W_n. This approach facilitates the adaptive enhancement of features at each level, thereby promoting precise saliency detection in computer vision applications.

B. Triple Feature Encoder Module

Traditional feature pyramid networks introduce a top-down path to generate multi-scale feature maps by upsampling high-level feature maps and fusing them with low-level feature maps. However, due to the insufficient interaction of semantic information between levels, it is difficult to effectively synergize the low-level detail information with the high-level semantic information, which affects the characterization ability of the fused feature graph. This paper proposes the Triple Feature Encoder Module (TFE) approach, which fuses three scales of feature information to generate high-quality semantic information. The design can not only enhance the characterization ability of features but also improve and refine the feature information.

Fig. 3 shows the structure of the TFE module. Here, C represents the number of channels and R the resolution of feature maps. L1, M1, and S1 denote the large, medium, and small feature maps from the output of the ATW module. The upsample uses nearest neighbor interpolation. For large-size feature maps (L1), a hybrid structure of maximum pooling and average pooling is utilized for down sampling, which is beneficial to preserve the validity and diversity of high-resolution features and small objects. Medium-size feature maps (M1) can be subjected to a convolution operation or without any untransformation. For small-size features (S1), the nearest neighbor interpolation method is used to adjust the resolution to 1R. The three changed features are then subjected to a concat operation, which undergoes a 1*1 convolution operation to modify the output channel. This approach helps to preserve the local feature richness of low-resolution images.

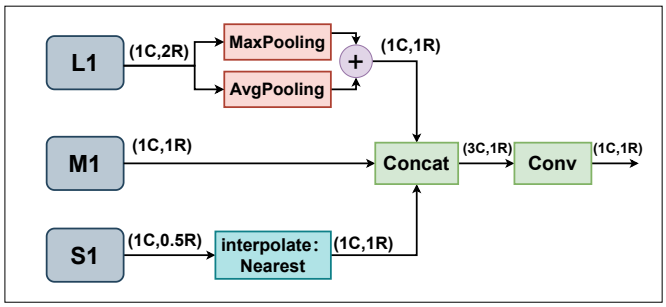

Fig. 3. The structure of the TFE module.

C. Global Attention Mechanism

However, small objects occupy fewer pixels and contain less information in the image, making them more susceptible to being ignored or misclassified during detection. The attention mechanism guides the network to prioritize the features of small objects, thereby enhancing their distinguishability in subsequent processing by improving the representation of their features. By facilitating the capture of long-range dependencies, this mechanism leverages context from surrounding pixels and the broader image, thereby augmenting the network's feature representation capabilities. However, both SENet and CBAM approaches overlook the interactions between channels and spatial dimensions, leading to the loss of cross-dimensional information. The global attention mechanism (GAM) [33] mitigates information loss and amplifies interactions across global dimensions. This enhancement bolsters the features of small objects, mitigates information loss, and thereby elevates the detection performance for such objects.

Fig. 4 shows the overview of the Global Attention Mechanism (GAM), where diagram A represents the overall input-output flow of the GAM, subdiagram B represents the flow of the channel attention mechanism, and subdiagram C represents the flow of spatial attention. Where F_1, F_2 and F_3 represent the input feature map, intermediate state map, and output feature map, respectively. The expressions are as follows:

$$F_2 = M_c(F_1) \otimes F_1 \,. \tag{3}$$

$$F_3 = M_s(F_2) \otimes F_2 \,. \tag{4}$$

where M_c denotes the channel attention mechanism, M_s denotes the spatial attention mechanism, and $\otimes$ denotes element-by-element multiplication.

The channel attention submodule arranges spatial information into 1 dimension and realigns dimensional positions. It subsequently applies a two-layer multi-layer perceptron (MLP) to enhance the interdimensional dependencies between channels and spatial features.

In order to expand the receptive field, the spatial attention mechanism uses 7*7 convolutional layers. To reduce the computational effort, the number of channels is regulated using the channel reduction rate r. Finally, the feature map is passed through a sigmoid activation function, which generates attention weights that indicate the degree of importance of different locations or features.

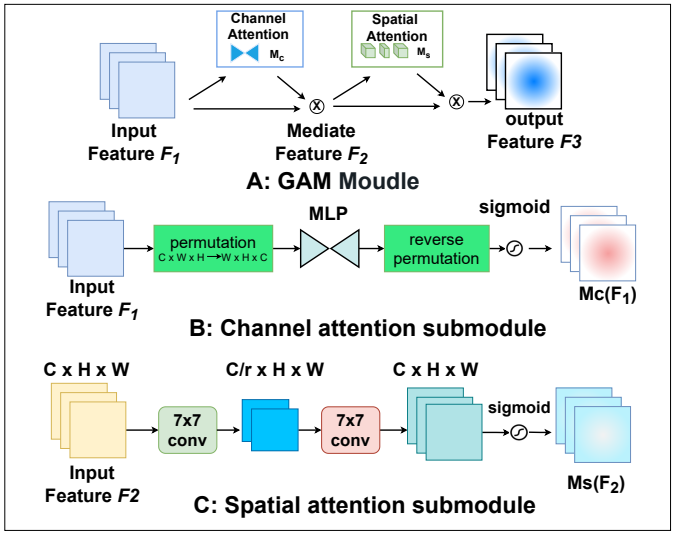

Fig. 4. The overview of the global attention mechanism.

IV. Experiments

To verify the effectiveness of the ATG-Net for small object detection, this paper conducts extensive experiments on the VisDrone2020 [34] dataset, a popular and challenging benchmark for aerial image detection.

1) VisDrone2020: This dataset contains a total of 10,209 images, of which 6471 were used for training, 548 for validation, 1610 for general testing, and 1580 for challenging testing. The image resolution of the dataset is approximately 2000×1500. The dataset encompasses 2.6 million annotations across various categories, primarily focusing on vehicles such as cars, buses, bicycles, tricycles, motorcycles, awning-tricycles, trucks, and vans, along with pedestrians, all captured from drone-based observations. It has extreme category imbalance and scale imbalance, making it an ideal benchmark for studying small object detection problems.

2) Implementation details: RetinaNet (Retina) [25], Faster R-CNN (FRCNN) [29], and Cascade RCNN (CRCNN) [39] are respective representatives of one-stage detectors, two-stage detectors, and cascade detectors. Accordingly, the paper designates them as the baseline detection networks for comparison. For data augmentation, the paper utilizes simple yet effective methods such as random resizing, random cropping, and random flipping. We implement the ATG-Net based on mmdetection on a single Nvidia 3060Ti GPU with 16GB of graphics memory. The optimizer employed is Stochastic Gradient Descent (SGD), initialized with a learning rate of 0.01. The learning rate strategy integrates both linear and cosine annealing schedules, initially employing a linear decay over the first 10 epochs, followed by a cosine decay for the subsequent 20 epochs, thereby encompassing a total training duration of 30 epochs. To assess the network's performance, Average Precision (AP) is utilized as the key metric. $AP_{50:95}$ is the average accuracy calculated over a range of different Intersection over Union (IoU) thresholds. It provides a more comprehensive picture of the model's performance under different IoU thresholds. AP_{50} and AP_{75} are computed at single

TABLE I. DETECTION RESULTS OF DIFFERENT NETWORKS ON THE VISDRONE2020 VALIDATION SET

Method	Backbone	$AP_{50:95}$	AP_{50}	AP_{75}	PED	PER	BC	Car	Van	Truck	TRI	ATRI	Bus	MO
Retina [26]	R50	13.9	27.7	12.7	13.0	7.9	1.4	45.5	19.9	11.5	6.3	4.2	17.8	11.8
FRCNN [26]	R18	21.8	39.2	21.5	18.1	12.9	7.3	50.3	30.5	21.5	15.5	8.1	34.8	18.7
FRCNN [26]	R50	21.7	39.8	21.0	21.4	15.6	6.7	51.7	29.5	19.0	13.1	7.7	31.4	20.7
FRCNN [26]	R101	21.8	40.2	20.9	20.9	14.8	7.3	51.0	29.7	19.5	14.0	8.8	30.5	21.2
CRCNN [26]	R50	23.2	40.7	23.1	22.2	14.8	7.6	54.6	31.5	21.6	14.8	8.6	34.9	21.4
FRCNN+ MMF [35]	R50	22.6	41.7	21.6	21.6	15.3	9.6	51.5	28.5	20.4	15.9	7.5	33.7	21.6
FRCNN+SimCal [36]	R50	20.0	35.8	19.6	18.7	13.8	5.7	51.0	28.4	16.4	13.6	5.9	27.0	19.4
FRCNN+RS+BGS [37]	R50	23.0	43.0	22.0	21.8	16.0	8.1	51.8	31.1	19.8	15.0	8.4	36.1	21.5
FRCNN+DSHNet [38]	R50	24.6	44.4	24.1	22.5	16.5	10.1	52.8	32.6	22.1	17.5	8.8	39.5	23.7
Retina+ATG-Net	R50	18.1	30.6	18.7	13.8	7.7	5.0	48.2	24.7	21.1	10.5	5.5	31.9	12.6
CRCNN+ATG-Net	R50	24.9	40.8	26.3	20.5	12.5	10.2	54.3	34.6	27.4	17.7	11.2	40.2	20.7
FRCNN+ATG-Net	R18	27.2	44.8	28.8	22.5	16.2	12.5	54.9	38.2	27.7	20.4	13.1	42.8	23.6
FRCNN+ATG-Net	R50	**28.9**	**46.8**	**30.9**	**23.7**	**17.2**	**13.8**	**56.2**	**39.7**	**30.4**	**22.6**	**13.9**	**47.1**	**24.9**

IoU thresholds of 0.5 and 0.75 across all categories. AP_s, AP_m and AP_l presents the average precision of the model in detecting small, medium, and large sizes receptively.

A. Experimentation Results

1) Comparison with baseline models: To demonstrate the effectiveness of the ATG-Net algorithm for detecting various types of targets on UAV images, the paper compares the proposed model with three baseline models and various improved FPN methods. The baseline models include Faster RCNN (FRCNN), RetinaNet (Retina), and Cascade RCNN (CRCNN), all evaluated under the same experimental conditions. ResNet18 (R18) and ResNet50 (R50) were chosen as the backbone networks. The evaluation metric for the object category utilizes $AP_{50:95}$. Experimental results with the baseline model and various improved FPN methods are shown in Table I. Where PED stands for pedestrian, PER stands for person, BC stands for bicycle, TRI stands for tricycle, ATRI stands for awning-tricycle, and MO stands for motor.

From Table I, ATG-Net achieves consistent performance improvements across all the detection networks with which it is combined. For Faster R-CNN, this paper uses three backbone architectures for comparative experiments. Notably, the R50 backbone yields the most significant performance boost, enhancing the $AP_{50:95}$ from 21.7% to 28.9%, representing a 7.2% improvement. When compared to the Retina model, there was an AP improvement from 13.9% to 18.4%, marking a 4.5% enhancement. The optimal detection model, Cascade R-CNN, likewise exhibits performance enhancement, with the AP advancing 23.2% to 24.4%. Upon incorporating our proposed ATG-Net module, all three baseline models experienced a significant improvement in detection accuracy across all categories. Notably, in categories like 'bicycle' and 'bus', which are underrepresented in the training data and typically appear very small, our method—employing FRCNN with the R50 backbone—achieves remarkable $AP_{50:95}$ increases of 7.1% and 15.7%, respectively. This highlights the ATG-Net's capability to excel at detecting small objects even when trained on limited data, affirming its robustness in such challenging scenarios.

Table I also presents the detection results of various advanced FPN networks improved upon FRCNN. ATG-Net also achieved the highest average detection precision, surpassing other detectors. In the detection of ten categories, ATG-Net has achieved good results, especially in the category of the bicycle and bus, where it outperforms DSHNet by 3.7% and 7.6%, respectively.

To further demonstrate the effectiveness of the ATG-Net model in detecting small objects, Table II is provided. A comparative analysis of various advanced object detection algorithms on the VisDrone2020 test set is presented. Combining ATG-Net with FRCNN and utilizing R50 as the backbone network, the optimal result was achieved on AP_{50}, with 38.4%. As shown in Table II, categories with a higher proportion of small targets, such as bicycles and buses, exhibit a substantial improvement, with the AP_{50} increasing to 18.2% and 64.4%, respectively.

B. Ablation Experiments

To validate the individual contributions of ATG-Net's feature pyramid components—ATW, TFE, and GAM—to the detection performance, ablation experiments were conducted. Experiments were conducted on the VisDrone2020 validation set using FRCNN as the baseline model and R18 as the backbone network. Table III shows the effect of each component of the ATG-Net on the detection performance.

1) Impact of TFE module: As shown in Table III, the addition of the TFE module increases AP_{50} from 39.0% to 42.4%. The AP_s increases from 14.1% to 16.6%, indicating that the TFE module can effectively improve the precision of small objects. This indicates that the TFE module can well fuse different levels of feature maps, which in turn enhances the model's ability to deal with multi-scale features, enabling the model to obtain better performance in the recognition of small and large objects.

2) Impact of the ATW module: The adaptive triple feature weighting module is able to adaptively predict a set of weights based on the importance of the triple features. ATW and TFE need to be used together. From Table III, when ATW and TFE are fused, AP_s increases from 14.1% to 17.3%. Combining the two modules enhances the model's robustness in detecting small targets. This also demonstrates that the ATW module effectively predicts weights from features of different scales.

3) Impact of GAM module: Although the use of GAM alone did not significantly improve detection performance, combining it with the other two modules enhanced the model's overall object detection capabilities. Compared to the baseline model, AP_{50} improves from 39.0% to 44.8%, an increase of 5.8%. For small objects, the AP increased from 14.1% to

TABLE II. COMPARISON OF EXPERIMENT RESULTS WITH OTHER POPULAR ALGORITHMS ON THE VISDRONE2020 TEST SET

Method	Backbone	AP_{50}	PED	PER	BC	Car	Van	Truck	TRI	ATRI	Bus	MO
CenterNet [40]	R50	26.6	22.6	20.6	14.6	59.7	24.0	21.3	20.1	17.4	37.9	23.7
YOLOv4 [41]	CSPDarknet53	32.5	28.2	15.9	5.8	65.7	25.2	26.1	13.8	8.1	40.2	26.1
YOLOv3-LITE [42]	DarkNet-53	28.5	**34.5**	**23.4**	7.9	70.8	31.3	21.9	15.3	6.2	40.9	32.7
MSA-YOLO [26]	CSPDarknet53	34.7	33.4	17.3	11.2	**76.8**	41.5	41.4	14.8	18.4	60.9	31.0
DINO [43]	Transformer	24.8	15.6	9.4	10.0	47.7	31.1	30.1	17.3	16.8	45.0	17.6
FRCNN+ATG-Net	R18	34.9	26.8	14.4	16.9	72.4	47.8	46.5	24.5	22.3	63.6	31.5
FRCNN+ATG-Net	R50	**38.4**	27.9	15.9	**18.2**	73.7	**50.4**	**49.1**	**24.8**	**25.0**	**64.4**	**34.8**

TABLE III. ABLATION STUDY RESULTS OF THE THREE COMPONENTS OF THE ATG-NET ON VISDRONE2020 VALIDATION SET. ✓ INDICATES THE USE OF THE MODULE

TFE	ATW	GAW	AP_{50}	AP_s	AP_l	param(M)
×	×	×	39.0	14.1	29.0	**121**
✓	×	×	42.4	16.6	33.9	110
×	×	✓	36.1	13.5	30.8	146
✓	✓	×	42.8	17.3	36.1	114
✓	✓	✓	**44.8**	**18.5**	**37.2**	162

18.5%, indicating that the model has excellent small object detection capability.

C. Visualization

In order to more intuitively demonstrate the effectiveness of the proposed method in practical application, some representative images from the Visdrone2020 test challenge dataset were selected for testing. All experiments were conducted by comparing the baseline FRCNN model, using R18 as the backbone network, with the model that combines our proposed ATG-Net with FRCNN.

Fig. 5 compares the visualization results of the highest-resolution feature map generated by the neck network. From left to right, the first column represents the original images, the second column shows the visualization results of the baseline model, and the third column displays the visualization results of our proposed ATG-Net. From the visualization results, it is evident that the feature maps produced by the baseline Faster R-CNN have a limited receptive field. This limitation suggests that the baseline model may struggle to capture detailed information or context over larger areas, leading to inaccuracies in detection. In contrast, the feature maps obtained by our ATG-Net have a global receptive field and focus on relatively smaller regions of interest compared to features of the same level. This characteristic allows our model to capture more detailed information and maintain context across different scales, thereby improving detection precision.

Fig. 6 shows the detect results on representative and more difficult images from the Visdrone2020 test challenge dataset. In this figure, the different categories are represented by different colored boxes, and the numbers on the rectangles indicate the confidence scores. The left column shows the results from the baseline FRCNN model with an R18 backbone. The right column shows the results from both FRCNN and ATG-Net models utilizing the same R18 backbone. Different categories in the detection results are identified using different colored detection boxes. Yellow boxes are used to highlight the detection of small objects, and zoomed-in effects are shown alongside for a more intuitive comparison. From the detection results, it can be seen that the baseline model has misdetections and misses small objects in the presence of occlusion, whereas the proposed model shows no misses and detects more small objects even in the presence of occlusion. In the detection effect image taken from high altitude, the vehicles and pedestrians on the road are very small. In this situation, the model in this paper can also detect them well. In the images of different lighting scenes, the model still has good detection ability in the dim scene.

V. CONCLUSION

In this paper, we proposed ATG-Net, an improved feature pyramid network for boosting UAV aerial image object detection. Firstly, we propose an Adaptive Triple Weighting (ATW) module, which intelligently assigns weights to predictions across diverse scales—large, medium, and small—dynamically emphasizing the significance of each size category. Secondly, we introduce a Triple Feature Encoding (TFE) module to utilize more efficiently on multi-scale contextual information. By applying the derived weights to features across various scales, this module amplifies and integrates multi-resolution features, thereby enhancing the representational capability of small object features. Due to the global attention mechanism (GAM) taking into account global information, it is crucial for enhancing the detection performance of small objects. Extensive experimental results on the VisDrone2020 have demonstrated that ATG-Net can effectively replace existing FPN networks and integrate with various popular detectors. Meanwhile, the proposed model can significantly enhance feature fusion capabilities, thus improving the detection precision of small objects. To enhance ATG-Net's detection capabilities even further, our next goal is to reduce model complexity and build lightweight detection models that can be deployed into edge devices.

REFERENCES

[1] Z. Li, Y. Zhang, H. Wu, S. Suzuki, A. Namiki, and W. Wang, "Design and application of a uav autonomous inspection system for high-voltage power transmission lines," *Remote Sensing*, vol. 15, no. 3, p. 865, 2023.

[2] A. Bouguettaya, H. Zarzour, A. Kechida, and A. M. Taberkit, "A survey on deep learning-based identification of plant and crop diseases from uav-based aerial images," *Cluster Computing*, vol. 26, no. 2, pp. 1297–1317, 2023.

[3] A. Utsav, A. Abhishek, P. Suraj, and R. K. Badhai, "An iot based uav network for military applications," in *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2021, pp. 122–125.

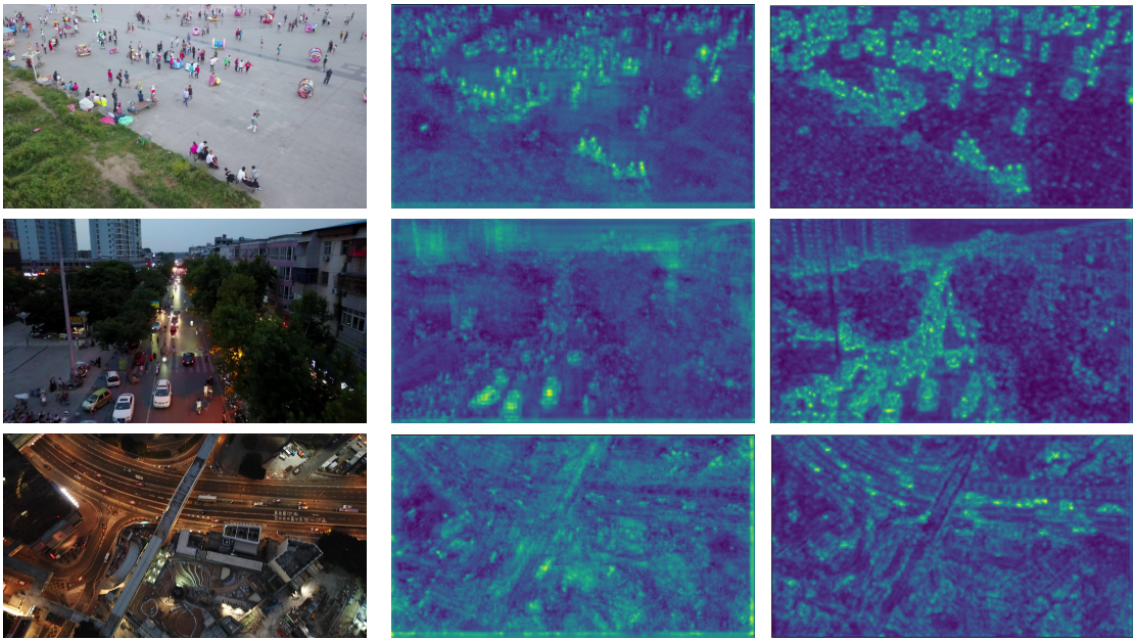

Fig. 5. Feature visualization results. From left to right, the column shows the original input image, the visualization result of the baseline model, and the visualization result of ATG-Net.

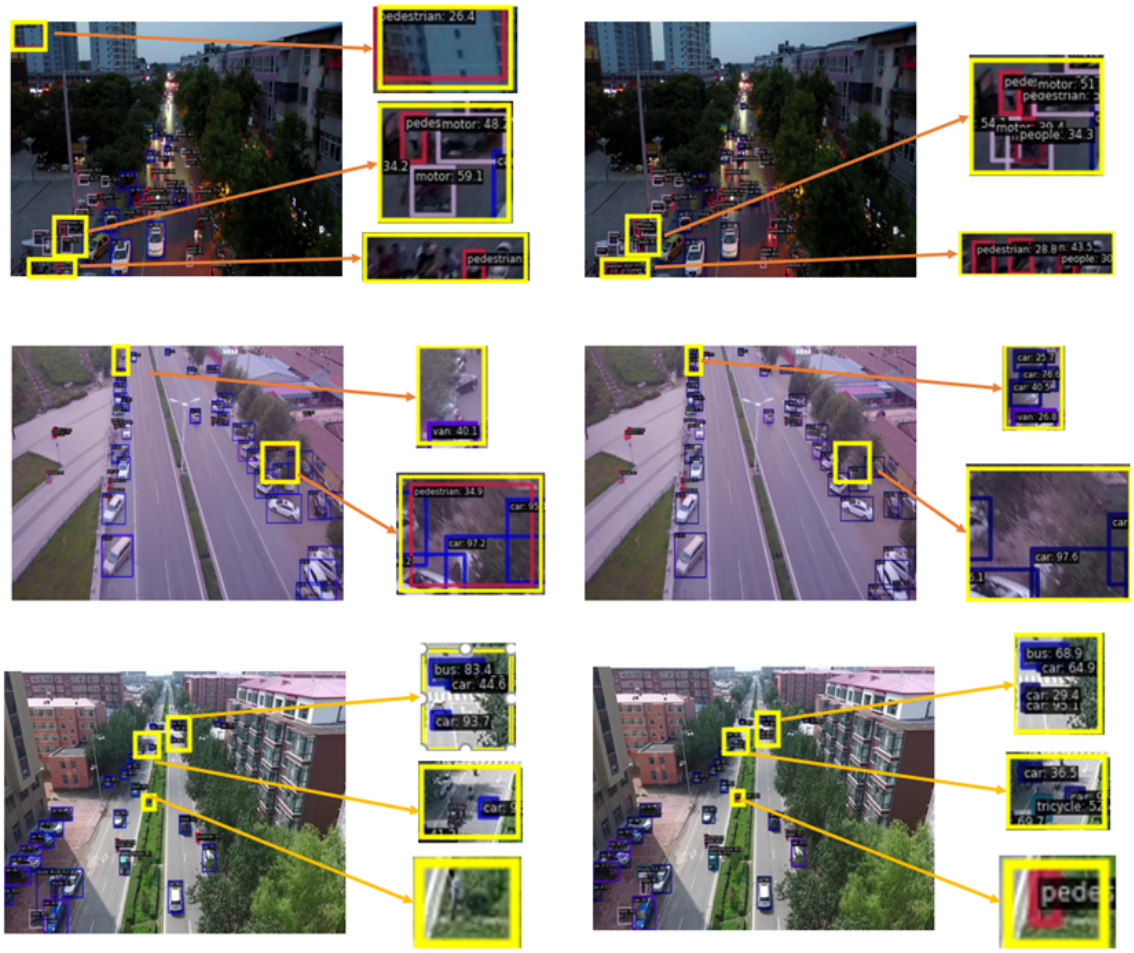

Fig. 6. Visualization results of challenging images on the VisDrone2020 validation dataset.

[4] W. Guettala, A. Sayah, L. Kahloul, and A. Tibermacine, "Real time human detection by unmanned aerial vehicles," in *2022 International Symposium on iNnovative Informatics of Biskra (ISNIB)*. IEEE, 2022, pp. 1–6.

[5] C.-J. Lin and J.-Y. Jhang, "Intelligent traffic-monitoring system based on yolo and convolutional fuzzy neural networks," *IEEE Access*, vol. 10, pp. 14 120–14 133, 2022.

[6] M. S. Mia, A. A. B. Voban, A. B. H. Arnob, A. Naim, M. K. Ahmed, and M. S. Islam, "Danet: Enhancing small object detection through an efficient deformable attention network," in *2023 International Conference on the Cognitive Computing and Complex Data (ICCD)*. IEEE, 2023, pp. 51–62.

[7] Y. Mo, J. Huang, and G. Qian, "Deep learning approach to uav detection and classification by using compressively sensed rf signal," *Sensors*, vol. 22, no. 8, p. 3072, 2022.

[8] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang *et al.*, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 454–14 463.

[9] G. Wang, Y. Chen, P. An, H. Hong, J. Hu, and T. Huang, "Uav-yolov8: a small-object-detection model based on improved yolov8 for uav aerial photography scenarios," *Sensors*, vol. 23, no. 16, p. 7190, 2023.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[14] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[15] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.

[16] 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," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[19] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[20] Y. Zhang, J. H. Han, Y. W. Kwon, and Y. S. Moon, "A new architecture of feature pyramid network for object detection," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. IEEE, 2020, pp. 1224–1228.

[21] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1968–1979, 2021.

[22] X. Zhou and L. Zhang, "Sa-fpn: An effective feature pyramid network for crowded human detection," *Applied Intelligence*, vol. 52, no. 11, pp. 12 556–12 568, 2022.

[23] Z. Li, C. Lang, J. H. Liew, Y. Li, Q. Hou, and J. Feng, "Cross-layer feature pyramid network for salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 4587–4598, 2021.

[24] L. Zhu, F. Lee, J. Cai, H. Yu, and Q. Chen, "An improved feature pyramid network for object detection," *Neurocomputing*, vol. 483, pp. 127–139, 2022.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[26] S. Zeng, W. Yang, Y. Jiao, L. Geng, and X. Chen, "Sca-yolo: A new small object detection model for uav images," *The Visual Computer*, vol. 40, no. 3, pp. 1787–1803, 2024.

[27] M. Kang, C.-M. Ting, F. F. Ting, and R. C.-W. Phan, "Asf-yolo: A novel yolo model with attentional scale sequence fusion for cell instance segmentation," *Image and Vision Computing*, vol. 147, p. 105057, 2024.

[28] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[30] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, and Y. Fu, "Rethinking classification and localization for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 186–10 195.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[33] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," *arXiv preprint arXiv:2112.05561*, 2021.

[34] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

[35] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and small object detection in uav vision based on cascade network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 118–126.

[36] T. Wang, Y. Li, B. Kang, J. Li, J. Liew, S. Tang, S. Hoi, and J. Feng, "The devil is in classification: A simple framework for long-tail instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 728–744.

[37] Y. Li, T. Wang, B. Kang, S. Tang, C. Wang, J. Li, and J. Feng, "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 991–11 000.

[38] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in uav images for object detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3258–3267.

[39] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[40] B. M. Albaba and S. Ozer, "Synet: An ensemble network for object detection in uav images," in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 10 227–10 234.

[41] S. Ali, A. Siddique, H. F. Ateş, and B. K. Güntürk, "Improved yolov4 for aerial object detection," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2021, pp. 1–4.

[42] H. Zhao, Y. Zhou, L. Zhang, Y. Peng, X. Hu, H. Peng, and X. Cai, "Mixed yolov3-lite: A lightweight real-time object detection method," *Sensors*, vol. 20, no. 7, p. 1861, 2020.

[43] N. D. Vo, N. Le, G. Ngo, D. Doan, D. Le, and K. Nguyen, "Transformer-based end-to-end object detection in aerial images," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01410113