Decoupled Feature Pyramid Learning for Multi-Scale Object Detection in Low-Altitude Remote Sensing Images

Haokai Sun, Yaxiong Chen, Xiongbo Lu, and Shengwu Xiong

***Abstract*—Recently, low-altitude remote sensing platforms are widely used for various practical applications. Object detection is a basic and significant technology, serving them. The scale imbalance problem is predominant in low-altitude remote sensing images, which brings a great challenge to detect objects from these imageries. Consequently, in this article, we boost performance from the perspective of mitigating scale imbalance issues. First, we choose a one-stage object detector with decoupled heads as the baseline because of its comparatively high efficiency and accuracy. Current-decoupled heads ignore the interlayer relationship and the information contained. On the other hand, all existing feature pyramid structures generate one feature map for two branches at every layer. Inspired by them, we propose a novel feature pyramid network paradigm—decoupled feature pyramid network with consideration of different preferences for classification and localization. Meanwhile, the introduction of feature pyramid architecture will cause performance deterioration of larger objects because upper layers receive insufficient supervision in the training phase. Therefore, we adopt a distinct supervision strategy—level supervision, which pays more attention to upper layers. We demonstrate extensive experiments on two popular benchmarks of object detection in low-altitude remote sensing images to validate the effectiveness of our proposed method. In addition, we introduce a scale imbalance metric to quantify the degree of size change of objects to better illustrate the ability to relieve the scale imbalance problem. Finally, our proposed approach achieves state-of-the-art performance on both datasets.**

***Index Terms*—Feature pyramid, level supervision, low-altitude remote sensing images, object detection, scale imbalance (SI).**

I. Introduction

NOWADAYS, low-altitude remote sensing platforms with characteristics of easy operation, low cost, and ability of real-time image acquisition are increasingly employed for numerous high-frequency practical applications, such as power inspection, traffic monitoring, and disaster rescue. Object detection is one of the significant and fundamental technologies of this wide range of applications. In object detection tasks, recently, with the development of deep neural networks, researchers have achieved satisfactory performance on public benchmarks such as MS COCO [1] and PASCAL VOC [2]. In consideration of the great success of object detection in generic scenarios, increased researchers adopt methods based on deep learning (convolutional neural networks and transformers) to detect the object in remote sensing images [3], [4], such as [5], [6], [7], [8].

However, due to particularities of low-altitude remote sensing platforms [9], the images captured by them differ from generic scenarios and bring huge challenges. Briefly, there are the following three primary difficulties of object detection in low-altitude remote sensing images.

1) Proportion of small objects is high and they distribute densely.

2) Computing resource is constrained but low latency is demanded.

3) Scale of objects is imbalanced.

In the past, many works delved into the former two questions and have attained fruitful results. Exploring the third problem more specifically, the agile flying altitudes of low-altitude remote sensing platforms cause the distances between the photography platform and objects to change sharply. Meanwhile, low-altitude remote sensing platforms have multiple viewpoints that make images include objects the near in larger and the far in smaller at the same time [3]. They both lead to the severe scale imbalance (SI) problem. We select two pictures with different flying altitudes and camera angles from the VisDrone dataset

Manuscript received 30 April 2023; revised 16 June 2023; accepted 24 June 2023. Date of publication 7 July 2023; date of current version 24 July 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160604, in part by the Project of Sanya Yazhou Bay Science and Technology City under Grant SCKJ-JYRC-2022-17 and Grant SCKJ-JYRC-2022-76, in part by the NSFC under Grant 62101393 and Grant 62176194, in part by the Major Project of IoV under Grant 2020AAA001, in part by the Youth Fund Project of Hainan Natural Science Foundation under Grant 6220N344, in part by the CAAI-Huawei MindSpore Open Fund under Grant CAAIXSJLJJ-2022-001A, in part by the Knowledge Innovation Program of Wuhan-Basic Research, in part by the Sanya Science and Education Innovation Park of Wuhan University of Technology under Grant 2022KF0020, in part by the Fundamental Research Funds for the Central Universities under Grant WUT:223110001, in part by the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX1148, in part by the Open Project of Wuhan University of Technology Chongqing Research Institute under Grant ZL2021-6, in part by the Hainan Special PhD Scientific Research Foundation of Sanya Yazhou Bay Science and Technology City under Grant HSPHDSRF-2023-03-002, and in part by MindSpore, which is a new deep learning computing framework (https://www.mindspore.cn/). *(Corresponding author: Shengwu Xiong.)*

Haokai Sun and Xiongbo Lu are with the School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China (e-mail: haokaisun@whut.edu.cn; luxiongbo01@gmail.com).

Yaxiong Chen and Shengwu Xiong are with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China, also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China, also with the Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China, and also with the Chongqing Research Institute, Wuhan University of Technology, Chongqing 401122, China (e-mail: 593544199@qq.com; xiongsw@whut.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3293395

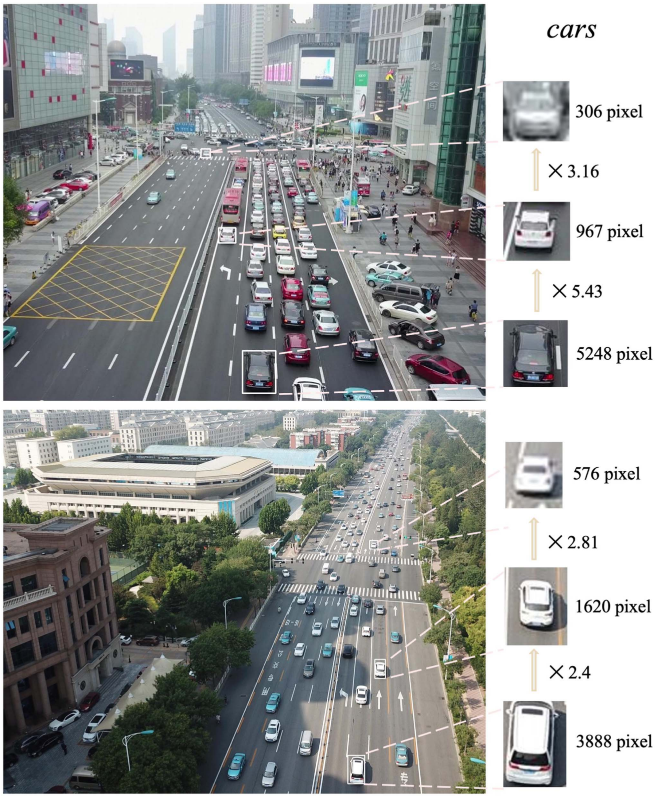

Fig. 1. Visualization of SI problem in low-altitude remote sensing images from the VisDrone dataset.

to display this issue vividly in Fig. 1. In these pictures, we can observe that sizes of objects vary dramatically even though they all belong to the same category—car and small instances are predominant.

In the field of generic object detection, employing cross-scale features is an effective way to mitigate this issue. Generating multiscale features from different stages of the backbone becomes the most popular way. FPN [10] builds a feature pyramid by integrating the features from lower and higher layers via a top-down pathway. It has achieved great success and has dominated modern detectors. After that several works [11], [12] follow FPN and attempt to gain more effective feature representations by trying various multiscale feature fusion strategies.

In the field of object detection in remote sensing images, the FPN architecture is also widely adopted to achieve better multiscale object detection. CAD-Net [13] introduces a spatial-and-scale-aware attention module to pay more attention to the regions with rich information and combine the global and focal information to attain more reliable features for objects in remote sensing images. FMSSD [14] proposes a spatial feature pyramid to leverage the information from multiscale and same scale feature maps. ABNet [15] designs an adaptive balanced network with an attention mechanism to gain more discriminative features. But their computation costs are too heavy to afford for low-altitude remote sensing platforms. Meanwhile, all existed variants of the feature pyramid follow the setting that each stage outputs one feature to the corresponding detection head, and then, the detection head processes the feature to extract useful information to complete the tasks of bounding box regression and classification. And increased attention mechanisms are introduced in remote sensing image tasks to promote the ability to discriminative extract information [16], [17]. Besides, several works enhance the multiscale feature extraction from the backbone. FSoD-Net [18] proposes a multiscale enhanced backbone to acquire more sufficient spatial and scale information. Lu et al. [19] united the advantages of convolutions and transformers to come up with a hybrid network for object detection in low-altitude remote sensing images.

As mentioned above, object detection in the low-altitude remote sensing platforms is expected to obtain certain detection accuracy while guaranteeing the higher detection speed. We consider that the one-stage detector pipeline without region proposal procedure, which has an instinct for a faster inference speed, is a more appropriate choice for low-altitude remote sensing platforms to achieve real-time object detection goals. Hence, our proposed object detector in this article adopts a one-stage framework. To further explore the composition of the one-stage object detector, we find that decoupled detection heads become the standard configuration of one-stage object detectors [20], [21] because they can partly alleviate the feature mismatch problem in object detection to gain better performance and speed up the overall training phase. Many researchers conducted works on it. But all previous works [22], [23], [24] produce specific features for two subtasks of object detection from the same feature input, which is the output from the feature pyramid without pondering interactions between layers in different scales.

In summary, SI is a prominent problem in low-altitude remote sensing imageries, which extremely harm the performance of object detection task. To deal with the SI problem in low-altitude remote sensing object detection, feature pyramids that utilize cross-features to create a more powerful feature representation at every stage with richer information is a universal and effective way. As all we know, with the help of decoupled heads in a one-stage detector, classification and bounding box regression tasks can obtain more precise corresponding features by a group of feature alignment operations to make a prediction improve the final detection accuracy. But they share the same feature and do not consider the influence of layer interaction. Based on them, we start to cogitate if we can design a novel feature pyramid paradigm, which can generate decoupled features for homologous decoupled heads. Concretely, this new decoupled feature pyramid network (DFPN) can generate more reliable feature representations for each decoupled head to better achieve two subtasks and, thus, improve the performance of the whole object detection with drastic changes in object sizes. Past works have already proved that classification and regression tasks have different preferences. The classification branch prefers the regions with richer semantic information than the regression one to infer the class of objects [23]. ASSD [25] notices that the feature misalignment problem also exists in remote sensing images for object detection and then constructs modules to align features. Therefore, we take different preferences of subtasks into account in the process of designing DFPN for remote sensing scenes.

On the contrary, the use of the FPN structure reduces the detection accuracy of some larger objects because of the supervision strategy lying in the training phase [23], [26]. The pyramid structure decides that the number of samples at the lower level is larger than at the upper level. We often adopt the method that calculates the training loss of different layers together and treats them fairly. It induces the shallow layer to receive more supervision than the deeper in the training phase. Hence, most of the experimental results exhibit that the utilization of the feature pyramid makes the detection of smaller objects better and larger objects worse. It can be regarded as a level imbalance problem. Although tiny instances account for the vast majority of low-altitude remote sensing images, there are still a great number of relatively larger objects. The proper detection of this part of objects is also something we cannot ignore in order that better deal with the problem of severe scale variation in low-altitude remote sensing images. We add level supervision for our object detector in the training progress to boost the attention on upper layers, making compensation for larger objects.

On the other hand, the quantitative measurement of the scale change of one category object is absent; thus, we introduce a calculation formula as metric to show the SI degree more clearly and directly. Then, combined with this metric, the effectiveness of our solution for the SI problem in low-altitude remote sensing images can be more fully illustrated.

In brief, the major contribution of our work lies in the following three aspects.

1) We propose a DFPN specifically designed for a one-stage object detector with decoupled heads adapted to the low-altitude remote sensing platform, alleviating the extremely prominent SI problem that existed in low-altitude remote sensing images.
2) We introduce level supervision in the training phase to compensate for objects in deeper stages with larger sizes, which further improves our model's ability to detect various scale objects in low-altitude remote sensing images.
3) We extensively demonstrate our method on two popular datasets of low-altitude remote sensing images—VisDrone [27] and UAVDT [28] to prove our proposed approach can mitigate the mentioned issue. Meanwhile, we utilize a metric to describe the degree of scale variation in one category to better illustrate the effectiveness of our means.

II. Related Works

A. General Object Detection

The current object detection methods based on deep learning can be mainly divided into one-stage or multistage detectors depending on whether they contain the procedure that generates a series of region proposals and then feeds them into another part to refine the prediction results. The multistage detectors contain Faster-RCNN [29], Mask-RCNN [30], and Cascade RCNN [31]. Faster-RCNN is a typical representative of multistage object detectors, which generate region proposals via a region proposal network. Mask R-CNN extends from Faster-RCNN, adding a branch to predict segmentation mask in parallel with object detection to acquire high-quality instance segmentation while taking object detection effectively. As for one-stage detectors, which omit the process of conducting region proposals, are represented by you only look once (YOLO) series [20], [21], [32], RetinaNet [33], FCOS [34], and GFL [35]. The YOLO family is famous for their efficiency and more advanced variants emerge endlessly. RetinaNet proposes a novel focal loss to address the class imbalance. FCOS is an anchor box-free and proposal-free detector, which avoids sophisticated computation.

In a word, the multistage object detectors usually hold higher detection accuracy than one-stage object detectors because of their refinement process, and meanwhile, their inference speed is relatively slower than one-stage detectors. Considering the characteristics of low-altitude remote sensing platforms, the computing resource is confined and insufficient but high efficiency is required, we incline to apply the one-stage pipeline in our work. And the decoupled heads are widely used in one-stage detectors because of their advantages [20], [22], [23]. They can mitigate the feature mismatch phenomenon in classification and localization to a certain extent. But they all share the same feature to generate corresponding task-specific features and neglect the important interaction between adjacent layers, losing some significant context information required. Their effectiveness of them is limited.

B. Object Detection in Remote Sensing Images

The object detection in low-altitude remote sensing images is different from, in general, pictures in the following three main aspects.

1) Smaller instances occupy the vast majority with dense distribution.
2) The higher effectiveness of balance between accuracy and efficiency is required.
3) The extreme scale variation of objects exists.

Most of the previous research works are based on the former two problems.

To deal with small object detection, they prefer to utilize a coarse-to-fine strategy [36], [37], [38], [39]. It designs a coarse detector that is responsible for locating large-scale targets and generating subregions that contain densely distributed small instances and a fine detector is responsible for further extracting small-size instances from these candidate regions. ClusDet [37] merges the object cluster and detection in one framework and the detection results come from fusing local and global predictions with NMS postprocessing. DMNet [38] proposes a new cropping strategy in aid of a density map. SB-MSN [36] uses multiscale feature pyramids and multistage heads to improve the quality of samples to train a better detector. Pipelines of these methods are like multistage detectors with more complicated architectures, there is no doubt that they cost a lot of time in the inference phase despite their comparatively higher accuracy. HRDNet [40] designs a multidepth and a multiscale feature pyramid network to enhance detection accuracy for small objects in high-resolution remote sensing images. Some works implemented multiscale inference or slicing-aided inference [19] to boost the performance at the expense of the amount of time consumed. Therefore, they are not appropriate ways for low-altitude remote sensing platforms to realize real-time object detection and cannot be employed in an actual production environment.

In view of hardware equipped on low-altitude remote sensing platforms being resource-constrained, several works have taken efforts in reaching a tradeoff between accuracy and efficiency. They leveraged sparse convolutions [41] to reduce computation costs and adopted a lightweighted structure [42]. QueryDet [43] uses sparse convolutions in the detection heads and creates new paradigms special for small objects. Recently, CEASC [44] proposes a novel plug–play detection head optimization approach based on context-enhanced sparse convolutions. Besides, Zhang et al. [45] adopted different strategies from the perspective of an automatic multiscale inference framework to trade off the balance between accuracy and efficiency.

C. Feature Pyramids

Feature pyramids have become one of the necessary components in nowadays object detectors, and they play an indispensable role in multiscale object detection. FPN [10] is a classic and landmark work, which utilizes a top-down pathway to fuse different scale features to obtain pyramidal representations with richer semantic information. After that, several variants explore more pathways to merge cross-features. PANet [11] adds an extra bottom-up pathway to enhance information signal transportation and a short-cut way to shorten information propagation, making the entire feature hierarchy able to fully use the lower layer information for accurate localization. M2Det [12] proposes a multilevel feature pyramid to produce multilevel, multiscale, and more representative features.

Many researchers in the remote sensing realm carried out research in view of multiscale object detection by improving the original FPN structures. Zhang et al. [46] used two feature pyramid networks, which are, respectively, responsible for region proposals and object detection to promote performance. CF2PN [24] adopts a cross-scale feature fusion method to generate multiscale features to mitigate SI. Sun et al. [47] proposed an end-to-end gated bidirectional network to eliminate interference information while fusing multiscale features. In addition, the FMSSD [14], ABNet [15], CAD-Net [13], and CANet [48] aim to enhance the ability of multiscale object detection in remote sensing images to a certain extent.

In general terms, the feature pyramid capably handles the difficulty of size change in object detection to some degree. However, all current works obey the rule that generates one feature map for the corresponding stage. The mainstream one-stage object detectors include decoupled heads as a standard configuration but decouple heads in every scale use the same original feature map to extract aligned features for classification and localization without ponderation of interlayer relationship and the information contained. Hence, we propose a decoupled feature pyramid especially for one-stage object detectors with decoupled heads to better deal with the SI problem in low-altitude remote sensing images in this article.

III. Proposed Method

A. Overview Framework

Compared with multistage object detectors, one-stage object detectors have a faster inference speed due to their simpler constructure without the region proposal procedure, which are more suitable for low-altitude remote sensing platforms with constrained computing resources to realize real-time object detection. Therefore, we select a one-stage detector—YOLOX [20] as our baseline, which owns better efficiency and accuracy. The entire network architecture of our object detector is shown in Fig. 2. Pondering the current feature alignment strategies in decoupled heads only based on the intralayer without interlayer consideration restrains the effectiveness. We introduce DFPN to replace the original feature pyramid to extract a more representative feature map for each branch and each layer, taking into the different preferences of classification and localization to learn more efficient multiscale information for the entire model. More details of it are discussed in Section III-B. In addition, we add level supervision on heads in the training phase to enhance supervision for some larger objects to make sure that different scale objects can be detected well to gain a higher accuracy overall. The specific method is displayed in Section III-C.

B. Decoupled Feature Pyramid Network

Multiscale object detection is a challenging task and vital for low-altitude remote sensing images with severe SI. In general, the object-detection realm implementation of feature pyramids has attained great success. The core idea of feature pyramid structures is fusing features from different stages—the lower to the upper, to extract the representative feature maps with effective multiscale information as more as possible. However, all current vanilla feature pyramid structures generate only one feature map for decoupled heads of corresponding layers and directly delegate the progress of aligning features for different subtasks—classification and localization to the decoupled heads. Unfortunately, these decoupled heads using one shared feature map to take alignment operation decide that they only align features from a spatial dimension without consideration of the interlayer relationship and joint optimization for two subtasks, which will compete. Hence, the effectiveness of improvement of performance is circumscribed. The DFPN, which is especially designed for one-stage detectors with decoupled heads emerges as the times require, is desired to adopt a more appropriate approach to fuse cross-layer features to elaborate respective features for different subtasks in the neck part. We hope the idea of decoupling used throughout the whole model design is to extract better multiscale information for different subtasks and to better solve the problem of SI in low-altitude remote sensing images.

The FPN is [10] a classical design in feature pyramids, widely used in modern object detectors. It uses a top-down pathway to merge the different stage features from the backbone. Recent

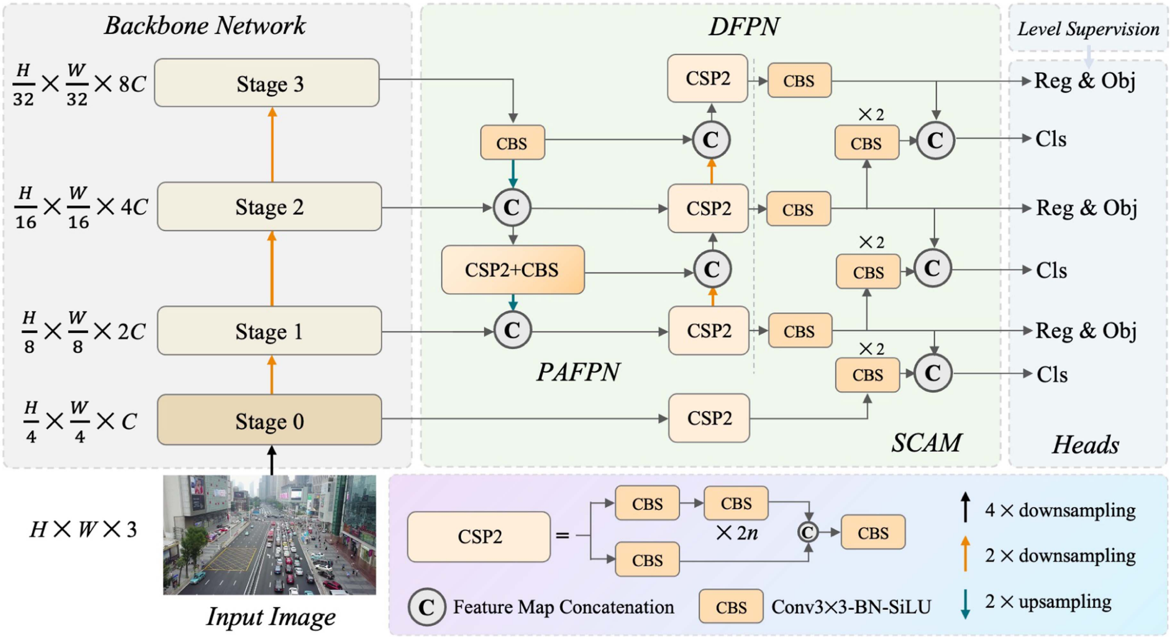

Fig. 2. Framework of our proposed method. DFPN consists of PAFPN and SCAM, generating feature maps separately for two branches of decoupled heads. The level supervision is introduced in the training phase.

works prefer to add an extra bottom-up way enlightened by PANet [11], finally integrating them to form a more powerful feature pyramid. We design our DFPN based on this strategy. Our baseline adopts three-layer PAFPN, we also follow this setting. The entire DFPN structure is shown in the middle of Fig. 2.

PAFPN consists of two pathways top-down and bottom-up and generates a series of feature maps $P = \{P_1, P_2, P_3\}$ with different scale $S = \{\frac{H}{8} \times \frac{W}{8}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{32} \times \frac{W}{32}\}$, where H and W are, respectively, behalf of the height and the width of input images. In past works, they are fed into heads directly. On the other hand, the features in the upper layer created by PAFPN with more channels are not necessary. It contains redundant information and cost enormous computation. Hence, we first use a CBS module to reduce the channel size of upper layers and, meanwhile, align the features from different scales that have the same channel size.

Past works [22], [25] have found that the classification and localization tasks have different feature preferences. Therefore, in the design of how to decouple the features from PAFPN, we put their preferences in the first place. Compared to the localization, the classification branch demands more semantic context information to distinguish the class of objects more accurately. And the localization task relies on the contour information to predict. Hence, after abundant experiments, for the localization branch, we choose the processed features from the CBS module as inputs P_l^{Loc} directly and where l represents the corresponding layer. Because the results of experiments show that features from PAFPN, which are suffered after sufficient multiscale information exchange and extraction are effective enough for localization without the need to utilize adjacent multilevel features to enhance. On the other hand, in the process of generating feature maps for the classification task, we introduce a semantic context augmented module (SCAM) to expand the semantic context information from multilevel features. The semantic information is important for classification, especially for some small objects, which dominate in low-altitude remote sensing images.

Semantic Context Augmented Module: For classification heads, we design SCAM to take full advantage of feature maps from two adjacent layers to obtain richer semantic information. Before fusing two-layer features, we implement downsampling operation to make sure they have the same scale. The above process of generating features for the classification branch can be written as follows:

$$P_l{}' = \text{Downsample}\left(\text{CBS}\left(P_l^{Loc}\right)\right) \tag{1}$$

$$P_l^{Cls} = \text{cat}(P_l^{Loc}, P_{l-1}{}') \tag{2}$$

where Downsample$(\cdot)$ is a CBS$(\cdot)$ module that stride is 2 and cat$(\cdot)$ means concatenate operation in channel dimension.

Besides, there is no output P_0 from PAFPN. Hence, we extract the feature map C_0 from Stage$_0$ of the backbone and make it through a CSP2 module like any other layers as $P_0{}'$. The CSP2 module is composed of two parallel branches and creates one output, which is a typical block in the YOLO family. The concrete design of its architecture is displayed at the bottom of Fig. 2.

Our DFPN generates task-specific features—P^{Cls} and P^{Loc}, corresponding to classification and localization branches. Different subtasks at different scales can obtain more matching features, which can help to better detect objects at different scales and achieve better multiscale object detection for these low-altitude remote sensing images.

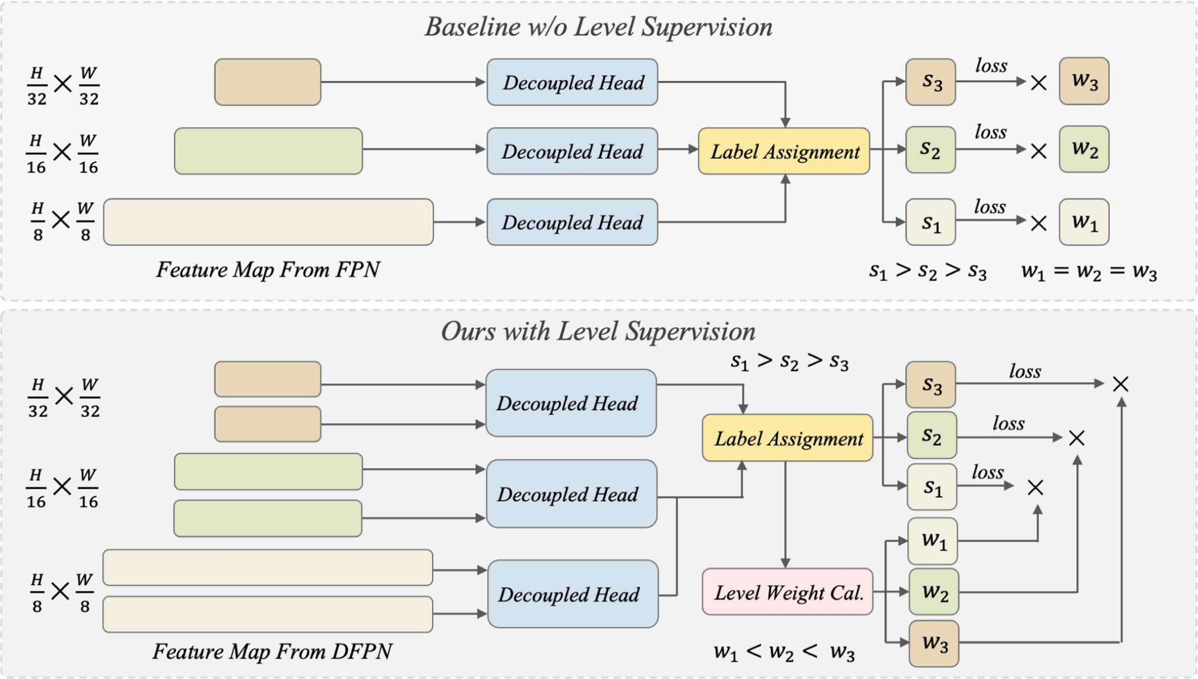

Fig. 3. Comparison between pipelines of the baseline without level supervision and ours with level supervision in the training phase. s_l and w_l denote the number of samples in layer l and the value of contribution weight in the entire loss for layer l, respectively.

C. Level Supervision

Indeed, the introduction of the feature pyramid network can realize better multiscale object detection to mitigate the SI issue in the low-altitude remote sensing images. On the opposite side, we can observe a phenomenon that the performance of respectively larger objects declines. Several works carried out further experiments and analysis to figure out the scheme of it. For example, the feature pyramid is revisited in view of optimization in YSLAO [26]. Their experiments show that the features generated by FPN for higher layers contain ineffective semantic information for larger objects and they believe that is why performance of larger objects is suppressed. Hence, they introduce the auxiliary loss to enhance the supervision of upper layers. QueryDet [43] has a similar view. In general, the upper layers, which are responsible for detecting a larger object, lack supervision.

To further explain it more explicitly, we draw a picture in Fig. 3. In the top part of the image is an illustration of the traditional approach. And the bottom part of the image specifically shows how we introduce level supervision into our detector. Takes our baseline with decoupled heads as an example, the feature pyramid network generates a series of different scale feature maps with scale after downsampling and upsampling operations in the entire process of fusion cross-layer features. The size of the feature map in a lower layer is twice as the higher layer. FPN feeds the processed multiscale features to corresponding heads and then the decoupled heads create predictions of classification and localization. In the training phase, then, we first combine all predictions from different layers and execute a certain strategy to assign all prediction sample labels, and all samples will be divided into positive and negative. These two kinds of samples will be treated differently in the progress of calculation loss. Finally, we handle the samples from different levels equally to calculate the loss. In other words, we give the same weight w_l value to all levels, the completed loss L can be formulated as

$$L = \sum_{l}^{n} w_l \left({L_l}^{Cls} + {L_l}^{Loc} \right) \tag{3}$$

where w_l denotes the weight of layer l. ${L_l}^{Cls}$ and ${L_l}^{Loc}$, respectively, represent the loss of classification and localization branches of layer l. n is the number of layers. In our baseline, n is 3 and w is $w_1 = w_2 = w_3 = 1$.

Obviously, the number of prediction samples in the lower layer is always greater than the upper. Because the sizes of feature maps from lower to upper layers are $\{\frac{H}{8} \times \frac{W}{8}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{32} \times \frac{W}{32}\}$ and one lattice in the feature map corresponds to one prediction produced, the numbers of samples generated from bottom to top of pyramid hierarchy are $\{\frac{H}{8} \times \frac{W}{8}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{32} \times \frac{W}{32}\}$, which express the lower layer holds four times as many samples as its contiguous higher layer. Hence, the loss L_l from the lower layer can contribute more to the whole L loss than the upper layer. In the process of backpropagation, the lower layers will receive more supervision. Based on the assumption that different layers detect objects from an exact size range from the former work, the smaller objects predicted by lower layers can obtain better performance than some larger objects predicted by upper layers. Furthermore, the overwhelming and unnecessary supervision of the smaller objects may deprive supervision of the larger objects, causing insufficient supervision for larger objects. It indicates that the former strategy that treats all samples fairly

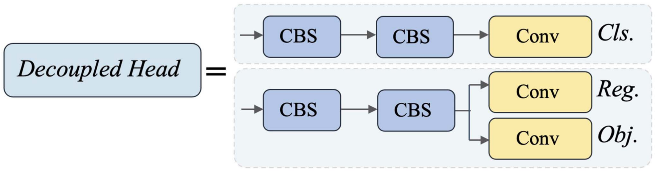

Fig. 4. Specific structure of decoupled heads.

from different layers without consideration of the instinct of the feature pyramid is not appropriate.

To compensate for the reduction accuracy of larger objects due to the introduction of the feature pyramid, we propose the level supervision to ensure that the upper layer can receive more supervision and place in the training phase. The diagrammatic drawing of our pipeline is shown at the bottom of Fig. 3. Compared to the current method, we introduce a level weight calculation module after the label assignment to create different weight w_l for different layers. The lowest level, which is responsible for smaller object detection with the largest number of candidate predictions, contributes most to the entire loss and receives the most powerful supervision signal, we give the weight a normal way and set $w_1 = 1$. The highest level needs more supervision, so we decided set $w_3 = 2$ to ensure that supervision signals are more likely to propagate to higher levels. The weight of the medium level dynamically adjusts according to the number of positive samples of the adjacent two layers to achieve better optimization results. Finally, considering the characteristic of distribution of samples in pyramid hierarchy, w_l are set as

$$w_l = \begin{cases} 1, & l = 1 \\ 2 - \frac{p_l - p_{l+1}}{p_{l-1} - p_{l+1}}, & l = 2 \\ 2, & l = 3 \end{cases} \tag{4}$$

where p_l is the number of positive samples assigned from layer l.

The structure of decoupled heads we utilize is exhibited in Fig. 4. For the localization head, the one in parallel with the bounding box regression branch is the object branch, responsible for predicting the probabilities of the foreground or background of the current position to facilitate more accurate localization. We follow the setting from our baseline, ${L_l}^{Cls}$ is computed by $Cross\ Entropy\ Loss$ and ${L_l}^{\text{Reg}}$ is calculated via IOU $Loss$ [49] for all positive samples, and ${L_l}^{\text{Obj}}$ also adopts $Cross\ Entropy\ Loss$ for all samples. After introducing level supervision, our whole $Loss$ is

$$L = \sum_{l}^{n} w_l \left({L_l}^{Cls} + {L_l}^{\text{Reg}} + {L_l}^{\text{Obj}}\right) \tag{5}$$

where ${L_l}^{\text{Reg}}$ represents the bounding box regression loss and ${L_l}^{\text{Obj}}$ means the loss of the object branch.

Besides, it is important to be emphasized that we utilize level supervision after certain epochs in the training progress. In a word, we should adopt a two-phase supervision strategy that treats samples from all levels equally at first and then introduce a level supervision strategy when lower layers responsible for smaller object detection receive necessary supervision signals. Otherwise, strengthening the supervision of larger objects too early will cause the decrement of detection performance of smaller objects, which dominate in low-altitude remote sensing images, and the paradoxical result is what we do not expect. Therefore, utilizing level supervision at the right moment is one of the key points to help us achieve more effective multiscale object detection in low-altitude remote sensing images with an SI issue.

IV. Experiments

A. Datasets and Metrics

We implement extensive experiments on two popular benchmarks for object detection in low-altitude remote sensing images to validate the effectiveness of our proposed method. One of them is the VisDrone dataset and another is the UAVDT dataset.

VisDrone dataset [27] collects 10 209 images with high resolution captured by low-altitude remote sensing platforms with different photography angles and various flying altitudes. It includes 6471 images used for training, 548 for validation, and 3190 for testing with 10 categories (pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor), across day and night. These images have a resolution of around 2000×1500. The former work adopts the validation dataset as the test dataset to evaluate the performance of approaches. Hence, all test results shown in the following contents are based on the validation dataset, consistent with former works.

UAVDT dataset [28] contains more low-altitude remote sensing images, compared to VisDrone. It is divided into the training dataset with 23 258 images and the testing dataset with 15 069 images with 3 kinds of common means of transportation (bus, truck, and car). All images are captured by cameras equipped on the low-altitude remote sensing platforms at low altitudes of urban regions. The resolutions of them are 1080×540.

Evaluation Metrics: We adopt AP, $AP50$, and $AP75$ three metrics used in MS COCO [1] to evaluate our proposed method just like previous works. We report them to validate the effectiveness of our object detector for low-altitude remote sensing images. The $AP50$ means that the average precision is obtained with an IOU threshold of 0.5. Likewise, the $AP75$ denotes that the set IOU threshold is 0.75 when calculating the precision. As for AP, it means the mean of a series of average precision with different IOU thresholds, from 0.5 to 0.95 at 0.05 intervals.

Besides, we introduce an SI metric to quantify the size change of objects of one category, and then, we can assess the SI degree more clearly and conveniently. Combined with this metric, the effectiveness of our DFPN in solving the SI problem, which is serious for low-altitude remote sensing images, can be more convincingly demonstrated. The SI metric SI are calculated as

$$\text{SI} = \frac{1}{\sum_{i}^{N} m_i} \sum_{i=1}^{N} \sum_{k=1}^{m_i} \frac{a_{i,k}}{a_{i,k-1}} \tag{6}$$

where N is the number of total images utilized and m_i represents the number of objects in ID i image. $a_{i,k}$ denotes the area of kth object in ID i image. A_i consists of all objects' areas of one category in ascending order, and we take $a_{i,1}$ value to $a_{i,0}$ in order to calculate conveniently. Hence, $A_i = \{a_0,\ a_1, \ldots, a_{m_i}\}$.

B. Implementation Details

We conduct our proposed method on PyTorch and MindSpore framework. We choose YOLOX-M [20] as our baseline and introduce a novel DFPN and level supervision for it to better deal with SI problems in low-altitude remote sensing images. The following shows the details of our implementation in the training phase and the testing phase.

Training phase: We use one NVIDIA A100 GPU to train our model on two benchmark datasets. For the VisDrone dataset, we set the initial learning rate to 0.0001 and adopt the StepLR learning rate scheduler with a multiplicative factor of learning rate decay of 0.92. The Adam optimizer is used where the weight decay is fixed as 0.0005. We train our model for 40 epochs in one period. We repeat it 4 times and the weights from the previous period will be loaded into the network before the next period starts. In general, our proposed methods are trained for 160 epochs. The level supervision is introduced after 160 epochs and we add an extra period to make sure it can improve the performance rather than counterproductive results. For the UAVDT dataset, all hyperparameters are fixed as same as for the VisDrone dataset. The difference is that we train models for two periods, and we utilize level supervision in the last period. For the sake of fairness, we decide not to adopt any data-augmented strategies such as copy–reduce–paste [50], Mosaic [32], and Mixup [51] on the experiments we conducted. In addition, the images are transformed to 640 × 640 before being transported to networks and the batch size is 4.

Testing phase: We evaluate the performance of models, using an NVIDIA RTX 3090 GPU. We infer one image once at a time. Similarly, to guarantee the fairness of reported results, we exclude any tricks used in the inference, including multiscales, and slicing-aided inferences [19]. The size of input images is set to 640 × 640 on both datasets. Past works prefer to use a relatively larger resolution of input images and regarding high-resolution images as inputs can boost the accuracy of performance. For a relatively fair comparison with state-of-the-art methods, we modify the resolution of images to 768 × 768 in Section IV-D.

C. Ablation Study

To support the effectiveness of our proposed DFPN and level supervision for object detection in low-altitude remote sensing images, we demonstrate broad experiments on both VisDrone and UAVDT datasets.

Decoupled Feature Pyramid Network: SI is predominant in low-altitude remote sensing images, which brings great challenges to object detectors for low-altitude remote sensing images. Thus, we come up with DFPN to extract more representative features with consideration of different preferences of classification and localization. And to explain more explicitly, we introduce SI as a metric to display the size change degree in a quantitative form. The SI is higher means that the size changes of this category are more dramatic. With the aid of SI, we can validate the effectiveness of DFPN more convectively. We test the $AP50$ metric of all 10 classes based on the PASCAL VOC [2] format. The results are shown in Table I and corresponding SIs of every category are reported in the second row of this table. With the help of DFPN, almost all of classes obtain improvement, except motor. Especially for the categories with higher SI, which means that SI is more severe, their performances increase more obviously. The $AP50$ of bus, van, truck, tricycle, and awning-tricycle with higher SI are promoted by 3.69%, 1.89%, 0.69%, 2.31%, and 1.19%, respectively. And the mean AP50 of all classes gains an 1.23% increment.

To further validate the generalization performance of our DFPN. We compare the AP, $AP50$, and $AP75$ with baseline using PAFPN on VisDrone and UAVDT, adopting the MS COCO protocol. All metrics acquire remarkable enhancement. As displayed in Tables II and III. For the VisDrone dataset, they are, respectively, improved by 0.8%, 1.1%, and 0.7%. For the UAVDT dataset, the values are 1.0%, 1.5%, and 1.3%. They all show the necessity of our DFPN to mitigate the SI issue in low-altitude remote sensing images.

Our DFPN is built based on PAFPN. We consider that the classification task is more sensitive to semantic information, so the SCAM module is added to generate features for classification especially. The idea of feature decoupling is used throughout the whole model to better solve the problem of feature mismatch of different tasks. Hence, the detection heads of each layer can better detect objects of different sizes. To further verify the correctness and effectiveness of the SCAM module and our decoupling feature pyramid idea, we separately conduct experiments on PAFPN and SCAM modules on the VisDrone dataset to analyze the contribution of each component to the overall performance. We set up two combinations—baseline+FPN and baseline+FPN+SCAM to eliminate the influence of the PAFPN module. The experimental results in Table II prove that adopting the mind of feature pyramid decoupling and strengthen semantic information extraction for the classification branch can effectively improve the accuracy of object detection in low-altitude remote sensing images. The AP, $AP50$, and $AP75$ are raised by 1.0%, 1.2%, and 1.1%, respectively. And for the PAFPN module, compared to the baseline with typical FPN, it can improve these metrics by 1.2%, 1.5%, and 1.5%.

Level Supervision: Although the introduction of feature pyramids can better realize multiscale objects detection, it also brings the deterioration of some larger objects because of the improper supervision strategy. Encouraged by this perspective, we utilize unfair supervision in the training phase, we call it level supervision. The main idea of it is that it gives different values for different layers in the calculation of the entire loss. On the other hand, we also consider that using level supervision too early will cause an enormous reduction of accuracy for smaller objects, which occupy the vast majority of instances in low-altitude remote sensing images. As a result, the overall performance will decrease rather than increase. It ought to be inappropriate. Therefore, we regard it as a fine-tuning strategy, and we introduce it to our model after several training epochs as described in Section IV-B. First, we conduct experiments on two benchmarks to confirm the ability of our level supervision for the

TABLE I
VALIDATION OF THE EFFECTIVENESS OF DFPN BY COMPARING $AP50$ OF EACH CATEGORY AND ANALYZING THE CORRESPONDING SI

Metric/Method	All classes	pedestrian	people	bicycle	car	van	truck	tricycle	awning-tricycle	bus	motor
SI	/	1.16	1.28	1.48	1.18	**1.92**	**2.41**	**1.96**	**2.00**	**2.54**	1.35
baseline	47.99	52.4	47.15	25.17	84.94	49.98	46.55	36.68	18.43	59.32	59.32
DFPN(w/o LS)	49.22	53.01	48.77	25.47	85.12	**51.87**	**47.24**	**38.99**	**19.62**	**63.01**	59.07

The bold values means that the category has higher SI.

TABLE II
ABLATION STUDY OF DFPN AND LS ON VISDRONE DATASET

Method	AP	AP50	AP75
baseline + FPN	26.1	45.7	25.8
baseline + FPN + SCAM	27.1	46.9	26.9
baseline + PAFPN	27.3	47.2	27.3
baseline + DFPN	28.1	48.3	28.0
baseline + DFPN + LS	28.3	48.7	28.3

TABLE III
ABLATION STUDY OF DFPN AND LS ON UAVDT DATASET

Method	AP	AP50	AP75
baseline + PAFPN	15.8	27.7	16.3
baseline + DFPN	16.8	29.2	17.6
baseline + DFPN + LS	17.1	29.3	18.1

TABLE IV
VALIDATION OF THE EFFECTIVENESS OF LEVEL SUPERVISION ON VISDRONE DATASET

Method	AP	AP50	AP75
baseline w/o LS	27.3	47.2	27.3
baseline w/ LS	27.8	48.0	27.8

object detector with DFPN. The results are reported in Tables II and III. We can observe that there are certain improvements in both VisDrone and UAVDT datasets after the implementation of level supervision. Indeed, the values of increment induced by LS are lower than DFPN. However, it does not influence the effectiveness of achieving more accurate object detection in low-altitude remote sensing images. And it proves that using LS can enhance the performance in a certain range, consistent with its own fine-tuning positioning. To further explore the capability of our level supervision, we compare the baseline with LS and without LS on VisDrone to excavate the potential ability, which is obscured by DFPN. We present the results in Table IV. The three metrics—AP, $AP50$, and $AP75$ are improved by 0.5%, 0.8%, and 0.5%, respectively. The data further verify the effectiveness of LS. And they confirm that LS can improve the performance to a greater extent when the initial accuracy is relatively low. In a word, it is a beneficial plug-and-play optimization approach without adding any extra computational cost in inferences.

Visualized results: We select two typical images from the validation dataset of VisDrone and visualize their prediction results from baseline and our proposed method in Fig. 5 to reflect the advancement of our approach intuitively. There are three columns, the first displays ground truths, the second exhibits the results predicted by the baseline, and the last shows the predictions from ours. We also magnify the regions with significant prediction differences and put them into gray squares under the corresponding images to compare their performance more distinctly. The above results indicate that our proposed method has a greater ability to deal with the SI problem. Our method can detect the small cars at a distance, which are omitted by the baseline. Additionally, in the second image, it can detect occluded objects, which are beyond the baseline's ability. And ours has higher classification accuracy and can detect tiny instances. In conclusion, our method obtains competitive performance in detecting multiscale objects in low-altitude remote sensing images.

D. Comparison With SOTA

We compare our proposed method with the state-of-the-art object detectors including one-stage detectors and multistage detectors. Several previous works prefer to adopt multiscale or slicing-aided inferences to enhance the performance. For ensuring fairness, we compare them without any tricks in inferences. On the other hand, the size of the backbone may influence the performance. Hence, other methods we reported utilize the backbones, which have a similar or greater magnitude than Modified CSP v5-M [20] adopted by us. Besides, we demonstrate experiments on both popular benchmarks to reflect the advancement of our object detector more fully.

For the VisDrone dataset, we select eight methods, containing RetinaNet [33] with ResNet50 [53], QueryDet [43] with ResNet-50, CEASC [44] based on GFL V1 [35] with ResNet-18, and some famous object detectors designed especially for low-altitude remote sensing images, including ClusDet [37], DMNet [38], and GLSAN [52] with ResNet-50, and HRDNet [40] with two backbones—ResNet-18 and ResNet-101. The results are shown in Table V. It is worth noting that our proposed method obtains the highest values across all three evaluation metrics, reaching the new state-of-the-art under comparable settings. For the UAVDT dataset, we also compare with ClusDet, DMNet, and GLSAN. In comparison with them, our DFPN with LS can boost the performance in AP and $AP50$ illustrated in Table VI. Particularly, the $AP50$ is increased by 1.5%. Those experimental results reveal that our model has attained state-of-the-art performance. More importantly, our object detector has tremendous potential for efficiency. The introduction of DFPN demands only slight extra cost time. And the usage of level supervision, which is only used in the training phase, will not cause any increment in inference times.

V. DISCUSSION AND CONCLUSION

There are three major challenges for object detection in low-altitude remote sensing images. We boost the performance from the perspective of multiscale object detection for low-altitude remote sensing platforms and consider the efficiency in this

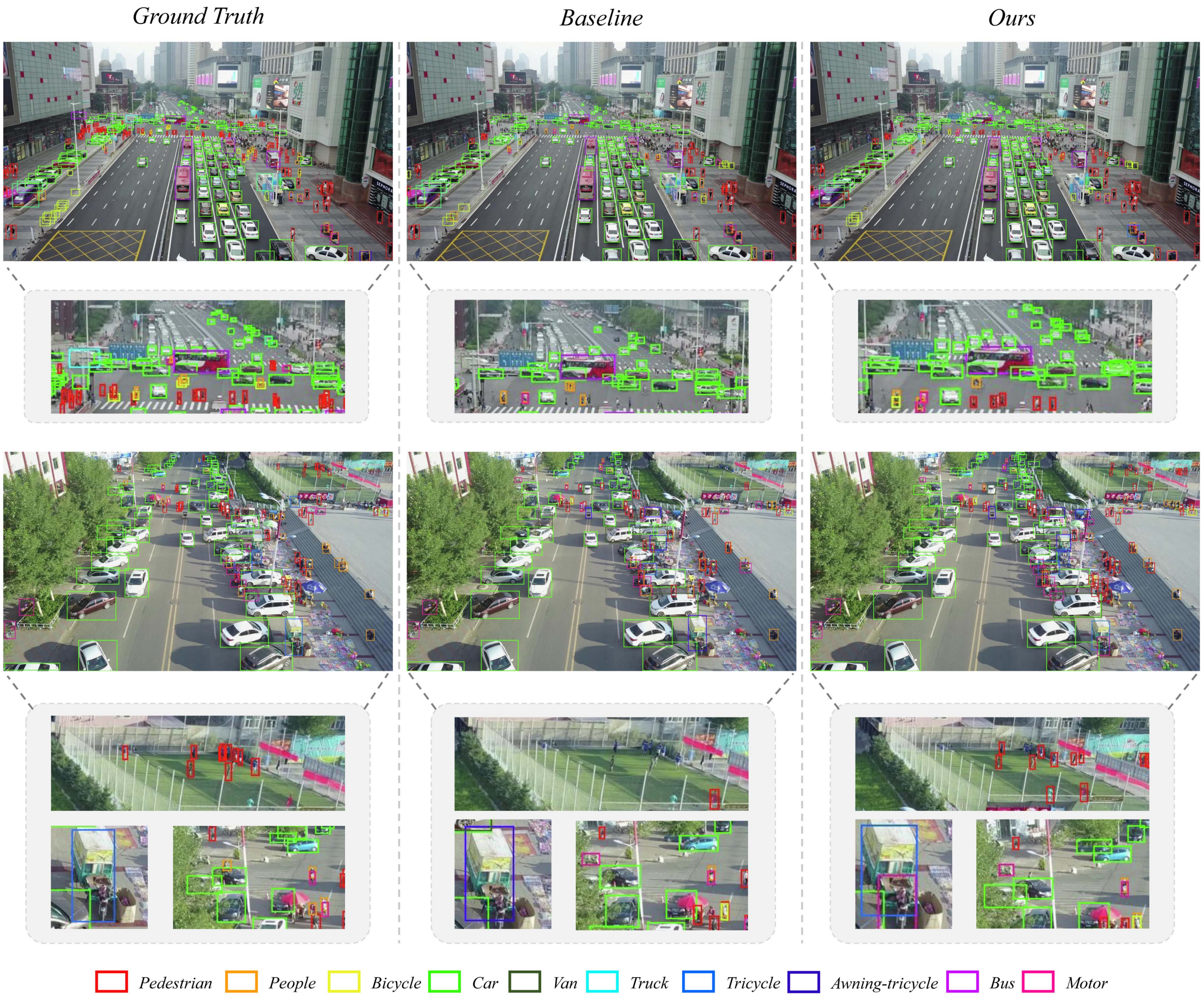

Fig. 5. Comparison between baseline and our proposed method by visualizing prediction results in two images from validation dataset of VisDrone. The contents in gray squares are magnified regions from corresponding images, which contain significant differences.

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON VISDRONE DATASET

Method	Backbone	Resolution	AP	AP50	AP75
RetinaNet [33]	ResNet-50	2400	26.2	44.9	27.1
ClusDet [37]	ResNet-50	1000 × 600	26.7	50.6	24.4
DMNet [38]	ResNet-50	1000 × 600	28.2	47.6	28.9
GLSAN [52]	ResNet-50	1000 × 600	25.8	51.5	22.9
HRDNet [40]	ResNet-18+ResNet-101	2666 × 1600	28.3	49.3	28.2
QueryDet [43]	ResNet-50	2400	28.3	48.1	28.8
GFL V1 (CEASC) [44]	ResNet-18	1333 × 800	28.7	50.7	28.4
DFPN(w/ LS)	Modified CSP v5-M	768 × 768	**30.3**	**51.9**	**30.5**

The bold values show the highest value in the corresponding column.

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON UAVDT DATASET

Method	Backbone	Resolution	AP	AP50	AP75
ClusDet [37]	ResNet-50	1000 × 600	13.7	26.5	12.5
DMNet [38]	ResNet-50	1000 × 600	14.7	24.6	16.3
GLSAN [52]	ResNet-50	1000 × 600	17.0	28.1	**18.8**
DFPN(w/ LS)	Modified CSP v5-M	640 × 640	**17.1**	**29.3**	18.1

The bold values show the highest value in the corresponding column.

article. At first, we choose a one-stage detector with decoupled heads as our baseline, with the consideration of its advanced accuracy and efficiency. We observe that the original decoupled heads neglect the effectiveness of the interlayer relationship and omit the information contained. The effectiveness of them is limited. Therefore, we design a novel feature pyramid network paradigm—DFPN to generate feature maps separately for classification branches and localization branches at every layer taking into account for different preferences of two subtasks. The introduction of the feature pyramid structure can mitigate the SI problem and bring the deterioration of the performance of larger objects at the same time. Second, although the ratio of smaller objects is so high, it should not be ignored to achieve better multiscale object detection. We utilize level supervision as a fine-tuning strategy to further enhance the accuracy of our model without any extra cost time in inferences. The core idea of our method is that boost the supervision for upper layers, which receive insufficient supervision in traditional ways to compensate for larger objects.

We conduct extensive experiments on two popular benchmarks for object detection in low-altitude remote sensing images—VisDrone and UAVDT to validate the effectiveness and advancement of our proposed method. The metric of quantifying the degree of the size change is absent; thus, we adopt an SI metric to describe the issue more clearly and more intuitively. With the aid of it, our proposed method can be better illustrated that have the ability to alleviate the SI in low-altitude remote sensing images and obtain greater accuracy. We also compare several SOTA methods with ours on both datasets, which indicates that our approach has reached advanced performance.

However, our proposed method still has some drawbacks and there are existing spaces for improvement. Although compared with other feature pyramid structures used to realize multiscale object detection in remote sensing images, the DFPN designed by us has a simpler structure and lower computational complexity; the overall design of ordinary convolution is still adopted, which limits the entire inference speed. And there is still a certain distance to meet the goal of real-time object detection on low-altitude remote sensing platforms. Therefore, in the future, we plan to make a lightweight design of the feature pyramid structure and consider adopting an architecture design based on sparse convolution, which is operated only over sparsely sampled regions or channels via learnable masks, thus limiting the amount of computation, to reduce the inference time to better trade off the balance between efficiency and accuracy.

References

[1] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.

[3] 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, 2020.

[4] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, “Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey,” *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 91–224, Mar. 2022.

[5] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, “Accurate object localization in remote sensing images based on convolutional neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[6] C. Wang et al., “Geospatial object detection via deconvolutional region proposal network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3014–3027, Aug. 2019.

[7] J. Xue, D. He, M. Liu, and Q. Shi, “Dual network structure with interweaved global-local feature hierarchy for transformer-based object detection in remote sensing image,” *IEEE J. Sel. Topics. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6856–6866, Aug. 2022.

[8] Q. Li, Y. Chen, and Y. Zeng, “Transformer with transfer CNN for remote-sensing-image object detection,” *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 984.

[9] Y. Chen, J. Huang, L. Mou, P. Jin, S. Xiong, and X. X. Zhu, “Deep saliency smoothing hashing for drone image retrieval,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 4700913, doi: 10.1109/TGRS.2023.3255302.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[11] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[12] Q. Zhao et al., “M2Det: A single-shot object detector based on multi-level feature pyramid network,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 9259–9266.

[13] G. Zhang, S. Lu, and W. Zhang, “CAD-Net: A context-aware detection network for objects in remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[14] P. Wang, X. Sun, W. Diao, and K. Fu, “FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[15] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, “ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5614914.

[16] X. Zheng, H. Sun, X. Lu, and W. Xie, “Rotation-invariant attention network for hyperspectral image classification,” *IEEE Trans. Image Process.*, vol. 31, pp. 4251–4265, May 2022.

[17] H. Sun, X. Zheng, X. Lu, and S. Wu, “Spectral–spatial attention network for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.

[18] G. Wang et al., “FSoD-Net: Full-scale object detection from optical remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5602918.

[19] W. Lu et al., “A CNN-transformer hybrid model based on CSWIN transformer for UAV image object detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1211–1231, Jan. 2023.

[20] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “YOLOX: Exceeding YOLO series in 2021,” 2021, *arXiv:2107.08430*.

[21] C. Li et al., “YOLOv6: A single-stage object detection framework for industrial applications,” 2022, *arXiv:2209.02976*.

[22] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, “TOOD: Task-aligned one-stage object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3490–3499.

[23] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, and F. Wu, “Disentangle your dense object detector,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4939–4948.

[24] W. Huang, G. Li, Q. Chen, M. Ju, and J. Qu, “CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection,” *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 847.

[25] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, “ASSD: Feature aligned single-shot detection for multiscale objects in aerial imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5607117.

[26] Z. Jin, D. Yu, L. Song, Z. Yuan, and L. Yu, “You should look at all objects,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 332–349.

[27] Y. Cao et al., “VisDrone-DET2021: The vision meets drone object detection challenge results,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2847–2854.

[28] D. Du et al., “The unmanned aerial vehicle benchmark: Object detection and tracking,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.

[29] S. Ren, K. He, R. B. 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.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[31] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[32] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[34] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[35] X. Li et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 21002–21012.

[36] W. Han et al., "Improving training instance quality in aerial image object detection with a sampling-balance-based multistage network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10575–10589, Dec. 2021.

[37] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8311–8320.

[38] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2020, pp. 190–191.

[39] O. C. Koyun, R. K. Keser, I. B. Akkaya, and B. U. Töreyin, "Focus-and-detect: A small object detection framework for aerial images,," *Signal Process: Image Commun.*, vol. 104, 2022, Art. no. 116675.

[40] Z. Liu, G. Gao, L. Sun, and Z. Fang, "HRDNet: High-resolution detection network for small objects," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[41] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, 2018, Art. no. 3337.

[42] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "R^2-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.

[43] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13668–13677.

[44] B. Du, Y. Huang, J. Chen, and D. Huang, "Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13 435–13 444.

[45] S. Zhang, X. Mu, G. Kou, and J. Zhao, "Object detection based on efficient multiscale auto-inference in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1650–1654, Sep. 2021.

[46] X. Zhang et al., "Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 755.

[47] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.

[48] L. Shi, L. Kuang, X. Xu, B. Pan, and Z. Shi, "CANet: Centerness-aware network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5603613.

[49] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 516–520.

[50] H. Ai, H. Zhang, and L. Ren, "An improved small target detection algorithm for SSD," in *Proc. Int. Conf. Mech. Robot.*, 2022, vol. 12331, pp. 50–58.

[51] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[52] S. Deng et al., "A global-local self-adaptive network for drone-view object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1556–1569, Dec. 2021.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

Haokai Sun received the B.S. degree in computer science and technology in 2021 from the Wuhan University of Technology, Wuhan, China, where he is currently working toward the Master of Engineering degree in computer science and technology with the School of Computer Science and Artificial Intelligence.

His research interests include machine learning and pattern recognition.

Yaxiong Chen received the B.Sc. degree in mathematics from Hubei University, Hubei, China, in 2014, the M.Sc. degree in mathematics from the Wuhan University of Technology, Wuhan, China, in 2017, and the Ph.D. degree in signal and information processing from University of Chinese Academy of Sciences, Beijing, China, in 2020.

He is currently an Associate Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China. His current research interests include pattern recognition, machine learning, hyperspectral image analysis, and medical imaging.

Xiongbo Lu received the B.S. degree in computer science and technology from the Hebei University of Technology, Tianjin, China, in 2014, and the M.S. degree in computer technology in 2019 from the Wuhan University of Technology, Wuhan, China, where he is currently working toward the Ph.D. degree in computer science and technology with the School of Computer Science and Artificial Intelligence.

His main research interests include scene text recognition and style transfer.

Shengwu Xiong received the B.Sc. degree in computational mathematics and the M.Sc. and Ph.D. degrees in computer software and theory from Wuhan University, Wuhan, China, in 1987, 1997, and 2003, respectively.

He is currently a Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan. His research interests include intelligent computing, machine learning, and pattern recognition.