Toward Efficient Object Detection in Aerial Images Using Extreme Scale Metric Learning

REN JIN1,2, JUNNING LV1, BIN LI1, JIANCHUAN YE1, AND DEFU LIN1
1Beijing Key Laboratory of UAV Autonomous Control, Beijing Institute of Technology, Beijing 100081, China
2School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
Corresponding author: Bin Li (3120150050@bit.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant U1613225w.

ABSTRACT In aerial image object detection, how to efficiently detect different size objects in input images of different scales and obtain a unified multi-scale representation of the object is an important issue. Existing methods rarely consider the connection between multi-scale training and multi-scale inference, and do not well optimize the constraint of input object samples in the multi-scale training process, which limits the performance of multi-scale representation. In this study, an efficient object detection algorithm for aerial images is proposed to alleviate this problem. Firstly, we propose to use metric learning to obtain the scale representation boundary of each object class, reduce the support of indistinguishable objects at extreme scales in the training process, and enhance the effect of multi-scale representation. Secondly, indistinguishable small objects are merged into small object regions, and these regions are trained to recommend the detector to detect small objects on the following high-resolution scale. Thus, a reasonable association between multi-scale training and inference is established, and the efficiency of multi-scale inference is considerably improved. The proposed algorithm has been tested on three popular aerial image datasets, including VisDrone, DOTA and UAVDT. Experimental results show that it can improve the detection accuracy and reduce the number of processing pixels.

INDEX TERMS Aerial images, extreme scale, metric learning, object detection.

I. INTRODUCTION

With the continuous increase in a large number of remote sensing equipment, automatic target recognition technology has gained significant importance. As a key algorithm, aerial image object detection has been widely used in target search [1], [2], multi-object tracking [3] and aerial surveillance [4], [5].

Aerial images usually have large image resolution (e.g., 2,000 × 1,500 resolution images in VisDrone [6]), large-scale changes (e.g., from 800 × 800 to 4,000 × 4,000 resolution in DOTA [7]) and many small crowded object instances, which cause difficulty in achieving the desired results in the direct use of popular deep learning-based object detection algorithm (e.g., RetinaNet [8], Faster RCNN [9], SSD [10], Yolo [11]). One of the important reasons is that the scale challenge leads to a decrease in aerial images’ representation learning performance compared to natural images. Singh et al. [12] proposed an efficient multi-scale training algorithm called SNIPER. This algorithm resamples low-resolution chips (512 × 512 pixels) on the image pyramid and improves training efficiency and multi-scale data augmentation. However, to avoid the adverse effects of extreme scale objects to the detector, SNIPER fixed the input object area range of each scale (e.g., for COCO [13] dataset, area ranges [0,802], [322,1502] and [1202,∞] are allowed for each of the scales). This approach leads to two problems: (1) for different datasets, the hyper-parameters of the area range need to be readjusted; (2) for different object categories, the threshold to distinguish whether the object is an extreme scale sample is also different. Figure 1 shows the histogram of the object size in VisDrone dataset. Multi-scale data augmentation brings abundant scale representation information and increases the minimum and maximum size of object samples. These samples that reduce detector performance are called extreme scale objects. Then, we discuss how to avoid the influence of these samples.

Figure 2 shows typical categories of normal samples and extreme scale samples in VisDrone dataset: (1) normal
In the stage of multi-scale inference, spending equal energy at every pixel in each scale is hugely inefficient. The problem becomes more serious in larger scale aerial images. Gao et al. [14] used reinforcement learning to determine the preferred search area in fixed-size grids. Najibi et al. [15] proposed a multi-scale inference algorithm combined with their SNIPER [12]. However, due to the use of fixed inference scales and the lack of extreme scale object estimation, their estimation of sub-region is not as accurate as of the proposed algorithm. The final detection accuracy also shows this point. Yang et al. [16] proposed a cluster region proposal network that detected dense object areas in aerial images to improve inference accuracy and speed; however, their available range is limited. For example, in DOTA [7] dataset, only dense objects, such as vehicles, can be addressed but not with targets, such as roundabouts and bridges. In this paper, multi-scale training and multi-scale inference are combined, in which multi-scale training enables the detector to recognize large and small objects on a single scale and recognize extreme scale objects in the next scale. The proposed method has no assumption of small dense objects and can effectively detect isolated objects in DOTA dataset.

In the aspect of multi-scale inference, an extremely small-scale region proposal network (ESRPN) and a scale estimation network are proposed to improve multi-scale detection performance and reduce the number of pixel processing.

In summary, the main contributions of this paper include the following:

1. A multi-scale training framework that uses the metric features of each class to eliminate indistinguishable objects in multi-scale data augmentation and improve the performance of multi-scale representation.
2. A multi-scale inference algorithm, including ESRPN and an improved scale estimation module, combined with small objects indistinguishable during training, forms small object regions and further infers on them, improves inference efficiency, and reduces the number of processed pixels.
3. On three representative aerial image datasets, namely VisDrone [6], DOTA [7], and UAVDT [17], we achieved the current state-of-the-art performance with less computation.

II. RELATED WORK
A. AERIAL IMAGE DETECTION
Compared with object detection in natural images, aerial image detection is more difficult because of various challenges, such as large scale, viewpoint change, many small crowded object instances, and limited airborne computing resources. This work focuses on detection that uses deep learning technology. We first review some studies on aerial image detection based on deep neural networks. Most studies focus on a specific target in aerial images or on rotation invariance, such as [18]–[21] and [22]. Sevo and Avramovic [23] introduced a two-stage CNN based detection algorithm into
FIGURE 4. Flowchart of the proposed extreme scale aware aerial image detection. In the training phase, the input image is scaled to multiple scales and clipped to uniform size chips. Then, metric learning is used to filter the extreme scale objects in rescaled ground-truth ROIs, and the extremely small objects are combined into the dense small object regions. These regions are ground-truth to train ESRPN. In the inference stage, the aerial image initially inputs the network at a lower resolution, detects the objects at this scale, and gives dense small object regions by ESRPN. Then, we cut out these small object regions on the original image and scale them to the appropriate inference scale by using optimal scale estimation network (OSEN) for detection. Finally, all the detection results are combined to provide the output.

aerial image object detection and achieved improved results. Audebert et al. [24] combined semantic segmentation with connected component object detection to improve performance. Sommer et al. [25] improved vehicle detection in aerial images by adjusting anchor size and increasing the resolution of the last feature map of Faster R-CNN [9]. Deng et al. [26] proposed a coupled region-based CNN to fuse more semantic information and improved detection accuracy. Ding et al. [27] proposed a region of interest (ROI) transformer to solve the rotation invariance problem in aerial images. Zhang et al. [28] proposed a scale adaptive proposal network that can consider large and small objects with a limited scale range. Yang et al. [16] proposed a clustered object detector to address the scale and sparsity challenges in aerial images. However, these methods do not analyze scale invariance in training and inference while scale invariance is an important problem in aerial images.

B. MULTI-SCALE TRAINING
Singh and Davis [29] analyzed the scale invariance of deep learning-based object detectors and obtained the following important conclusions: (1) Modern CNNs are not robust to changes in scale. (2) Scale invariant detector performs better than scale specific detectors because the former can capture more variation across the objects than the latter. (3) Training a detector using appropriately scaled objects in a certain input image scale is important. Singh et al. [12] proposed an efficient multi-scale training method called SNIPER. This method replaced every pixel’s input in an image pyramid and trained the detector with context regions around ground-truth instances cropped from the image pyramid, which is called chips. SNIPER can accelerate the training process and improve accuracy. However, both methods use a fixed pixel area threshold to filter the training objects. The filtered objects are called extreme scale objects. For aerial images with larger scale changes, more thresholds need to be searched, and the same threshold cannot fully describe objects of different categories.

C. MULTI-SCALE INFERENCE
Multi-scale inference can consider large and small targets. In aerial images with more scale changes, this inference can improve the detection accuracy effectively. The studies in multi-scale inference often focus on reducing the search area, thereby accelerating the processing. Lu et al. [30] proposed a method to detect the sub-regions with small and sparse objects adaptively. Alexe et al. [31] introduced a context-driven search method that can effectively localize the specific class object regions. Chen et al. [32] formulated the region search as a Markov decision process to learn contextual relations and refined the query class's search area. Gao et al. [14] used reinforcement learning to select and process valuable sub-regions on higher resolution. Lalonde et al. [33] proposed a two-stage spatial-temporal CNNs to detect vehicles in wide aerial motion imagery. Their key idea is to search the moving object regions. Yang et al. [16] proposed a clustered region search method and achieved improved efficiency in aerial images. However, these methods do not use the close relationship between the training scale and the inference scale. In this paper, extreme scale object learning is proposed,
which integrates training multi-scale and inference multi-scale, and can improve the detection accuracy and inference efficiency at the same time.

III. APPROACH

The main framework of the proposed algorithm is shown in Figure 4. Metric learning is used to distinguish and exclude extreme scale objects in the training process. Excluded extremely small-scale objects are combined into extremely small-scale regions to train ESRPN. In the inference stage, the extremely small-scale regions generated by the network are used to detect small objects in aerial images efficiently. Next, we introduce the selection of extreme scale objects.

A. EXTREME SCALE METRIC LEARNING

Chip is a cropped region in training images with a fixed size. It can not only sample more scale changes of input images better but also improve the utilization of GPU memory to achieve efficient multi-scale training. Also, according to the observation in [12], objects with extreme scales may degrade detection performance. SNIPER used a desired area range $R^i = [r^i_{\text{min}}, r^i_{\text{max}}], i \in \{1, n\}$ to determine which ground-truth boxes join in training for each $i$. For example, ground-truth object area from the ranges $[0, 80^2], [32^2, 150^2]$, and $[120^2, 1000^2]$ are used in training with the input scales 512, 800 and 1400, respectively. However, according to our observation, an apparent object can still be excluded from the training target on one scale. Providing a clear definition of extreme scale is challenging. Moreover, only using area range is not enough.

To solve the above problems, we propose to learn a metric space for extreme scale objects and use it to determine the detection network’s input data in the training phase. Inspired by recent work on deep metric learning [34], [35], we design a metric learning-based network to filter extreme scale objects. First, the following definitions are reviewed. Given a paired metric learning-based network to filter extreme scale objects.

Next, we use multi-scale input without constraints to train various models, such as Faster-RCNN with ResNet-50 [36], ResNet-101 [36], and ResNeXt-101 [37]. Given a small scale test image with resolution ranging from 320 to 800 with its short side, we label the objects that all models fail to detect as extreme scale objects. Then, we cut these objects from the small-scale image to train the network with metric learning. The sub-network structure of extreme scale metric learning is shown in Figure 5. The loss function of training extreme scale object classification network is detailed as follows:

$$L_{\text{cls}} = (1 - \gamma)L_{\text{softmax}} + \gamma L_{\text{metric}}$$

For a given ground-truth ROI, the subnet’s output embedding vector is represented as $E$. The probability of the given ROI in each class $i$ is formulated as:

$$p_i(E) \propto \exp\left(-\frac{d^2(E)}{2\alpha^2}\right)$$
where $d_i(E) = \min_j d(E, C^j_i)$ indicates the distance between embedding vector $E$ and clustered feature center $C^j_i$ of class $i$ and center $j$. The $i$th class distribution is assumed as mixtures of isotropic multivariate Gaussians with variance $\sigma^2_i$. In the stage of detection network training, for a ground-truth ROI, if $p_i(E) > T_p$ and the class $i$ belongs to extreme scale, then we terminate the training of this ROI. The experiment on parameter $T_p$ will be introduced in the section of the ablation study.

### B. ADAPTIVE ZOOM-IN WITH SMALL OBJECTS

For objects that are difficult to recognize in the pre-training model, we use metric learning to select extreme scale objects and classify them into three categories. The first category refers to the extremely large-scale object. The second one is the extremely small-scale object. The last one is the object with the annotation error or cutting operation that makes distinguishing difficult. In the inference stage, we mainly focus on small objects that are difficult to recognize at the current scale because they may be detected after zooming in on the input image. Inspired by the cluster proposal net of Yang et al. [16], we propose an ESRPN to generate the small object regions for detection on the next scale. The loss of ESRPN is expressed as follows.

$$L_{ESRPN}(\{p_i\}, \{t_i\}) = \frac{1}{N_B} \sum_i L_{BCE}(p_i, p_i^*) + \frac{\lambda}{N_l} \sum_i \mathbb{I}_{p_i^* > 0} I_{IOU}(t_i, t_i^*)$$  \hspace{1cm} (8)$$

where $L_{BCE}$ is the binary cross entropy, and $I_{IOU}$ is the IOU loss as in [38], IOU loss is used to increase the regression accuracy when anchor distribution is sparse. $p_i$ is the predicted probability of anchor $i$ being an extremely small-scale region, and $t_i$ is a parameterized vector of the predicted bounding box as same as the original RPN. $p_i^*$ and $t_i^*$ are ground-truth labels that correspond to anchor $i$, and $\mathbb{I}_{p_i^* > 0}$ is the indicator function, being 1 if $p_i^* > 0$, and 0 otherwise.

Next, we discuss the generation of ground-truth extremely small-scale regions to train the network. First, we provide each ground-truth object with a zoom-in value score.

$$s_g = \frac{p_i(E(g))}{g_w \cdot g_h}$$  \hspace{1cm} (9)$$

where $E(g)$ indicates the ground-truth object’s embedding vector obtained from the metric network; $p_i(E(g))$ is the probability that the ground-truth object becomes an extreme scale object; and $g_w$ and $g_h$ are the ground-truth box’s width and height, respectively. $\kappa = 0.1$ is used to offset the bias of small objects, and they can perform better after zoom-in.

Finally, a multi-scale sliding window is used to score each extremely small-scale region, and the criteria for scoring is described as:

$$ESS_b = (2 \sum_{g_i \in b} s_{g_i} - \sum_{g_i \in b} s_{g_i})/(b_w \cdot b_h)\nu$$  \hspace{1cm} (10)$$

where $b_i$ is an inner box in the extremely small-scale area, and $b_{in,x,y,1} = b_{cx,cy} - \xi b_{w,h}, b_{in,x,y,2} = b_{cx,cy} + \xi b_{w,h}, b_{in,x,y,1}$ and $b_{in,x,y,2}$ being the left/top and right/down coordinates of inner box, $b_{cx,cy}$ and $b_{w,h}$ are the central coordinate and width/height of ESS box, respectively; $\xi = 0.9$ is a scaling factor if the distance to the image boundary is more than $1 - \xi$ times its width/height; otherwise $\xi = 1.0$. $\nu = 0.5$ is used to balance the size of larger search bounding box with more small objects on average. The output extremely small-scale regions filter low score parts by threshold $T_x = 0.05$. Then, non-maximum suppression (NMS) algorithm is used to sparse the remaining regions, and finally the ground-truth of extremely small-scale regions is obtained. Some visual illustrations of extremely small-scale object scoring and clustering are shown in Figure 6.

### C. ADAPTIVE ZOOM-IN INFERENCE

In the process of inference, the test image is first downsampled to a lower resolution (short side length 480 in our experiments) and then fed into the feature pyramid network (FPN) to obtain the multilevel feature maps. On the feature maps, we apply two parallel region proposal networks, one is the traditional RPN network, which is used to propose object candidates, and the other is the ESRPN we mentioned above, which is used to extract extremely small-scale area and further detection after zooming. Next, these extremely small regions need scale estimation to obtain the optimal detection scale and are fed to the next scale detection network. Finally, we fuse all the sub-region detection results and provide the final detection results.

### D. OPTIMAL SCALE ESTIMATION

To obtain the optimal detection scale, we first evaluate the average precision (AP) of the detection network for object detection at different scales; the results are shown in Figure 7.

Take VisDrone [6] dataset as an example, and it can be seen that the objects have a relatively high mAP in the scale range
be defined as $L$, output scale, and relative scale, respectively. $s_i$ is the ground-truth relative scale from reference scale and $N$ is the number of extremely small-scale areas, and $smooth_{L1}$ is a smoothly $l_1$ loss function from Girshick [39].

IV. EXPERIMENTS

A. DATASETS

Our experiments are mainly carried out on three publicly accessible datasets: VisDrone [6], DOTA [7], and UAVDT [17].

1) VisDrone

This dataset consists of 263 videos, 179,264 frames, and 10,209 static images. These images are captured by various drone-mounted cameras in a wide range of aspects, such as different capture heights, diverse locations, environments, and object densities. For object detection task, VisDrone has 10,209 fully annotated images with 10 categories, where 6,471 images are used for training, 548 images for validation and 3,190 images for testing. The image scale of the dataset is approximately 2,000 × 1,500 pixels.

2) DOTA

This dataset is collected from multiple cities with multiple sensors and platforms (e.g., Google Earth), and its resolutions are from 800 × 800 to 4,000 × 4,000 pixels. The fully annotated images include 188,282 instances and 15 categories (e.g., large-vehicle, swimming pool, helicopter, bridge, plane, and ship). For the sake of comparison, we use the horizontal bounding box (HBB) detection task of DOTA to evaluate our proposed method. Consistent with the settings in [16], the images with movable objects in the dataset are selected to participate in the evaluation (e.g., small vehicle, large vehicle, plane, helicopter, and ship). The training and validation images are 920 and 285, respectively.

3) UAVDT

The UAVDT benchmark consists of 100 video sequences that are captured at various locations in the urban area. These sequences are taken on the unmanned aerial vehicle platform and contain a total length of more than 10 hours. Typical scenes include squares, main roads, toll stations, highways, intersections, and T-shaped intersections. These videos are recorded at a speed of 30 frames per second, and the image resolution is 1,080 × 540 pixels. A total of 23,258 training images and 15,069 test images are contained in this dataset.

B. IMPLEMENTATION DETAILS

We use ResNet50 and FPN [40] as our baseline models, which are based on the publicly available implementation maskrcnn-benchmark [41] and PyTorch. The efficient multi-scale training framework [12] is added to our training process, and the chip size is set to 512 × 512 pixels. We use three training scales (−1, 512), (800, 1200), and (1400, 2000), which are the default settings in [12]. Then, different models, including Faster-RCNN with ResNet-50 [36], ResNet-101 [36], and ResNeXt-101 [37], are trained using multi-scale input without any constraints. Extreme scale objects are defined as ground-truth that cannot be recalled by all models, where $IOU$ is 0.5 and score threshold is 0.2. Then, metric learning is used to classify normal and extreme scale objects, where $\alpha = 25$, $\beta = 0$, $m_1 = 1.2$, $m_2 = 0.4$, $\gamma = 0.6$, $T_p = 0.4$, and the feature extraction network is ResNet50. When training the proposed adaptive zoom-in model, the ground-truth bounding


1) EFFECT OF METRIC LEARNING

Metric learning is introduced to select extreme scale objectives. Intuitively, we have the following expectations: on the basis of correct detection of normal objects, extreme scale objects need to be selected as many as possible. Therefore, the influence of metric learning on the classification accuracy of extreme scale objects is compared. In the validation dataset, a serious imbalance between extreme scale objects and normal objects is observed. Moreover, the number of normal objects is larger than that of extreme scale objects. Thus, reflecting the average accuracy of extreme scale targets is difficult.

Here, the classification accuracy of extreme scale objects is evaluated at the cost of 5% normal objects after multi-scale data augment, where 5% is an empirical data, mainly from our ablation experiments on the SNIPER algorithm. In the default configuration of SNIPER, we use the minimum object area threshold to control the number of training samples. Then, the test performance is evaluated at a single scale (the short side of the input image is scaled to 800 pixels), and the results are shown in Table 2. It can be seen that when all the samples are involved in the training, AP is 29.7; when 1.2% of the minimum samples are filtered out, AP is 30.8; when 7.3% of the samples are filtered out, AP has begun to decline to 29.8; when 7.3% of the samples are filtered out, AP is lower than 29.8. Therefore, we choose 5% as the empirical parameter. The parameter \(p_r\) is adjusted to make the classification accuracy of normal samples greater than 0.95.

| Methods       | backbone         | training scale | filtering samples | test scale | AP    | AP_{0.5} | AP_{0.75} | AP_{m} | AP_{m+} |
|---------------|------------------|----------------|-------------------|------------|-------|----------|----------|--------|--------|
| FRCNN+FPN     | ResNet50         | 800            | -                 | 800        | 21.8  | 39.5     | 21.1     | 14.3   | 30.7   | 34.4   |
| FRCNN+FPN     | ResNet101        | 800            | -                 | 800        | 22.1  | 40.0     | 21.5     | 14.3   | 31.9   | 36.2   |
| FRCNN+FPN     | ResNeXt101       | 800            | -                 | 800        | 23.1  | 41.7     | 22.7     | 14.9   | 33.9   | 37.3   |
| FRCNN+FPN     | ResNet50 300-800 | 300-800        | -                 | 800        | 26.5  | 46.2     | 26.5     | 19.3   | 36.2   | 37.6   |
| FRCNN+FPN     | ResNet101 300-800| 300-800        | -                 | 800        | 26.8  | 46.4     | 26.6     | 19.5   | 36.5   | 38.4   |
| FRCNN+FPN     | ResNeXt101 300-800| 300-800        | -                 | 800        | 27.2  | 46.5     | 26.9     | 19.6   | 36.6   | 38.6   |
| FRCNN+FPN+SNIPER | ResNet50 512,800,1400 | -             | -                 | 800        | 29.7  | 50.7     | 30.5     | 20.4   | 42.0   | 51.0   |
| FRCNN+FPN+SNIPER | ResNet101 512,800,1400 | -             | -                 | 800        | 30.1  | 51.1     | 31.0     | 20.5   | 42.9   | 52.4   |
| FRCNN+FPN+SNIPER | ResNeXt101 512,800,1400 | -             | -                 | 800        | 30.5  | 51.2     | 31.2     | 20.7   | 43.2   | 52.8   |
| FRCNN+FPN+SNIPER | ResNet50 512,800,1400 [-1.80,][32,150,][120,-1] | -             | -                 | 800        | 23.4  | 38.7     | 24.9     | 13.0   | 41.8   | 35.5   |
| FRCNN+FPN+SNIPER | ResNet101 512,800,1400 [-1.80,][32,150,][120,-1] | -             | -                 | 800        | 23.4  | 38.8     | 24.7     | 12.5   | 42.2   | 36.9   |
| FRCNN+FPN+SNIPER | ResNeXt101 512,800,1400 [-1.80,][32,150,][120,-1] | -             | -                 | 800        | 23.5  | 42.1     | 23.1     | 15.1   | 34.3   | 37.0   |
| ESM(ours)     | ResNet50         | 512,800,1400   | extreme scale     | 800        | 32.2  | 56.3     | 33.8     | 21.5   | 44.3   | 52.5   |
| ESM(ours)     | ResNet101        | 512,800,1400   | extreme scale     | 800        | 33.6  | 57.8     | 34.6     | 23.0   | 45.7   | 52.7   |
| ESM(ours)     | ResNeXt101       | 512,800,1400   | extreme scale     | 800        | 33.9  | 58.4     | 34.7     | 23.2   | 45.9   | 53.2   |

C. EVALUATION METRICS

AP with different IoU thresholds is used as our main evaluation metrics, including AP at IoU thresholds 0.5, 0.75 or from 0.5 to 0.95 (AP_{0.5}, AP_{0.75}, AP_{0.5:0.75:0.95}). Other metrics include AP_s, AP_{m}, and AP_l, which indicate the AP for small objects with an area less than 32^2, medium objects with area from 32^2 to 96^2, and large objects with an area larger than 96^2, respectively. The details of these metrics are followed based on the evaluation protocol in MS COCO [13], VisDrone [6] and DOTA [7] (HBB task).

D. ABLATION STUDY

To validate the contributions of the extreme scale-aware detector and adaptive zoom-in inference, we conduct extensive experiments on VisDrone [6] dataset. To verify whether the method can achieve consistent performance improvement under different complexity backbone networks, three backbone networks, including ResNet-50 [36], ResNet-101 [36], and ResNeXt-101 [37], are used in the following experiments. In the test phase, the input image resolution is 800 pixels with its short edge for ESOM and ESOM-DZ, the input image resolution is 416 pixels for ESOM-DZ-416.
object or not; softmax+class is used to classify whether it is an extreme scale object and the correct class label simultaneously; softmax+class+metric is the result of introducing metric learning.

Table 3 shows that the loss of object class labels contributes to extreme scale objects’ classification accuracy. The metric loss can improve the accuracy of the classifier further.

Considering that metric learning involves many parameters, such as \( \alpha, \beta, m_1, m_2, \) and \( \gamma \), the ablation results of these parameters are shown in Figures 9 (a) to (e), respectively. \( \alpha = 25, \beta = 0, m_1 = 1.2, m_2 = 0.4, \) and \( \gamma = 0.6 \) are the best parameter combinations for extreme scale object classification in the VisDrone dataset. In the classification accuracy experiment, \( T_p \) can control the trade-off between the classification accuracy of normal samples and extreme scale samples. For example, the smaller \( T_p \), the higher segmentation accuracy of extreme scale samples, which will reduce the classification accuracy of normal samples; on the contrary, if \( T_p \) is high, then the higher detection accuracy of normal samples, which will increase the misjudgment of extreme scale samples. By adjusting \( T_p \), the detection accuracy in normal samples becomes above 0.95. That is, 5% of the number of normal objects is lost in multi-scale data augmentation to observe the classification accuracy of extreme scale samples. Therefore, in actual detection experiments, ablation experiments must be performed on \( T_p \) again to determine the \( T_p \) at the optimal detection precision. Based on VisDrone dataset and ResNet-50 backbone network, the results of the \( T_p \) ablation experiment for ESOM are shown in Figure 9 (f), where the green line represents the number of extreme scale samples filtered under the current \( T_p \). The blue line represents the average detection precision under the current \( T_p \), wherein \( T_p = 0.4 \) is the best choice on VisDrone dataset.

2) EFFECT OF MULTI-SCALE TRAINING
The experimental results of multi-scale training are shown in Table 1. As seen, multi-scale training can increase detection performance effectively. The prototype multi-scale training strategy can improve AP by 4%~5% compared with single-scale training. SNIPER allows the input of large-scale images and training with chips, which are beneficial to training small objects in aerial images. However, the extreme scale object will degrade the detection performance. Thus, it is an important issue to judge whether the object is on the extreme scale. SNIPER uses the bounding-box area of each object directly to divide them. The proposed method uses metric learning to classify whether the target is in the extreme scale and achieves better detection performance.

Besides, some local details of detection results under different multi-scale strategies are given in this section (Figure 10). Among the different strategies, the backbone network is ResNet50, and the visualized score threshold is 0.5. The proposed algorithm (e.g., ESOM and ESOM-DZ) in this paper is more effective for learning object scale-invariant features, such as tiny pedestrians in the first row, trucks in the second row, the large truck (bounding-box more accurate) in the third row, and the yellow car surrounded by pedestrians in the fourth row. Also, the DZ detection (ESOM-DZ) that will be introduced next can achieve better small object detection results (last column in Figure 10).

3) EFFECT OF ADAPTIVE ZOOM-IN
We compared the proposed adaptive zoom-in with single scale inference, multi-scale inference, and AutoFocus from [15]. The detection performance in Table 4 shows that our adaptive zoom-in only adds a few processing pixels to single scale inference and achieves performance similar to multi-scale inference. In addition, compared with AutoFocus [15], the proposed method has some improvements in terms of the number of pixels and in the accuracy. This finding is due to our extremely small-scale metric learning net.

To obtain better ground-truth values generated by extremely small-scale regions, we compared the effects of

| Models                  | Backbone | Percentage of SR | AP  |
|-------------------------|----------|------------------|-----|
| FRCNN+FPN+SNIPER        | ResNet50 | 0%               | 29.7|
| FRCNN+FPN+SNIPER        | ResNet50 | 1.2%             | 30.8|
| FRCNN+FPN+SNIPER        | ResNet50 | 4.1%             | 29.8|
| FRCNN+FPN+SNIPER        | ResNet50 | 7.3%             | 29.1|

TABLE 2. The results of ablation experiments on reducing the number of samples in multi-scale training on VisDrone dataset. SR is short for sample reduction.

| Classification loss function | Acc\(_{\text{extreme}}\) | Acc\(_{\text{normal}}\) | \( T_p \) |
|------------------------------|-------------------------|------------------------|----------|
| softmax                      | 0.532                   | 0.951                  | 0.56     |
| softmax+class                | 0.667                   | 0.952                  | 0.51     |
| softmax+class+metric         | 0.743                   | 0.951                  | 0.44     |

TABLE 3. The effects of different classification loss function on the accuracy of extreme scale objects.

FIGURE 9. Impact of different hyperparameters in metric learning on accuracy, where (a)-(e) are the effects of hyperparameters \( \alpha, \beta, m_1, m_2, \) and \( \gamma \) on the classification accuracy of extreme scale objects. (f) is the effect on the final detection accuracy under different \( T_p \).
different parameters on the recall rate and average search pixel area of small-scale targets (targets with an area less than \(16^2\) in VisDrone dataset) in the experiment. The results are shown in Figure 11. High-grade parameters should ensure the high recall rate of small objects while maintaining the small average search pixel area to ensure computational efficiency. As shown in Figure 11, \(\xi = 0.9\), \(\kappa = 0.10\), and \(\nu = 0.5\) are remarkable choices. Although the average search area is reduced to a reasonable level, it can still ensure a high recall rate for small objects. In practical use, the parameters can also be adjusted according to the requirements. For example, when the detection speed needs to be improved, the parameter combination with a lower search area can be selected.

4) EFFECT OF SCALE ESTIMATION

After obtaining the extremely small-scale regions, we need to perform inference on these regions in the image. Figure 7 shows that when the target has an area of \(16^2 \sim 64^2\) pixels in the image, the average detection accuracy is the highest. These extreme scale regions are then scaled in their object areas, which have an approximate value of \(32^2\) pixels. The ground-truth scale can be calculated from the average area of the ground-truth bounding box of objects in the region. Then, in this section, the proposed scale estimation network’s performance evaluation results are shown in Table 5. We use

![Figure 10. Comparison of local details of detection results under different multi-scale strategies shows that ESOM-DZ has significant advantages in addressing object scale changes and detecting small objects.](image-url)

![Figure 11.](image-url)
TABLE 7. The detection performance of the baselines and proposed method on the DOTA [7] dataset.

| Methods          | backbone | test pixels | AP  | AP_{50} | AP_{75} | AP_{s} | AP_{m} | AP_{l} |
|------------------|----------|-------------|-----|---------|---------|--------|--------|--------|
| RetinaNet [8]+FPN [40]+EIP | ResNet50  | 2440^2     | 24.9| 41.5    | 27.4    | 9.9    | 32.7   | 30.1   |
| RetinaNet [8]+FPN [40]+EIP | ResNet101 | 2140^2     | 27.1| 44.4    | 30.1    | 10.6   | 34.8   | 33.7   |
| FRCNN [9]+FPN [40]+EIP    | ResNet50  | 2410^2     | 31.0| 50.7    | 32.9    | 16.2   | 37.9   | 37.2   |
| FRCNN [9]+FPN [40]+EIP    | ResNet101 | 1490^2     | 31.5| 50.4    | 36.6    | 16.0   | 38.5   | 38.1   |
| ClusDet [16]          | ResNet50  | 1490^2     | 32.2| 47.6    | 39.2    | 16.6   | 32.0   | 50.0   |
| ClusDet [16]          | ResNet101 | 1490^2     | 31.6| 47.8    | 38.2    | 15.9   | 31.7   | 49.3   |
| ESOM-DZ (ours)        | ResNet50  | 1410^2     | 33.4| 52.4    | 37.1    | 16.9   | 38.0   | 48.6   |
| ESOM-DZ (ours)        | ResNet101 | 1390^2     | 33.7| 53.1    | 38.5    | 17.2   | 38.9   | 49.2   |

TABLE 8. The detection performance of the baselines and proposed method on the UAVDT [17] dataset.

| Methods          | backbone | test pixels | AP  | AP_{50} | AP_{75} | AP_{s} | AP_{m} | AP_{l} |
|------------------|----------|-------------|-----|---------|---------|--------|--------|--------|
| R-FCN [49]       | ResNet50  | 770^2      | 7.0 | 17.5    | 3.9     | 4.4    | 14.7   | 12.1   |
| SSD [10]         | VGG-16   | 512^2      | 9.3 | 21.4    | 6.7     | 7.1    | 17.1   | 12.0   |
| RON [50]         | N/A      | 384^2      | 5.0 | 15.9    | 1.7     | 2.9    | 12.7   | 11.2   |
| FRCNN [9]+FPN [40]| VGG-16   | 960^2      | 11.0| 23.4    | 8.4     | 8.1    | 20.2   | 26.5   |
| FRCNN [9]+FPN [40]+EIP | ResNet50 | 1920^2     | 6.6 | 16.8    | 3.4     | 5.2    | 13.0   | 17.2   |
| ClusDet [16]     | ResNet50  | 1000^2     | 13.7| 26.5    | 12.5    | 9.1    | 25.1   | 31.2   |
| ESOM-DZ (ours)   | ResNet50  | 980^2      | 14.1| 27.4    | 13.2    | 9.3    | 26.0   | 32.3   |

FIGURE 11. Effects of different parameters on the generation of extremely small-scale regions, where (a)∼(c) are the effects of hyperparameters $\xi$, $\kappa$, and $\nu$ on the tradeoff between recall rate and search pixel area.

the root mean square error (RMSE) as the evaluation criteria and compare it with ScaleNet in [16]. The proposed $1 \times 1$ and $3 \times 3$ convolution kernels can improve the performance of scale estimation effectively.

E. QUANTITATIVE RESULTS

1) VisDrone

The comparative experimental results on VisDrone are shown in Table 6. The algorithms involved in the comparison include popular common object detectors like Faster RCNN and RetinaNet, and small object friendly detectors such as RefineDet [42], CFE-SSD [43] and DPNet [44]. Besides, we also select some detectors for aerial images. Sommer et al. [25] improved the detection performance of small objects in aerial images by modifying the network structure and deploying data augmentation.

Vandersteegen et al. [45] optimized anchor size and tested different pre-training models to improve aerial images’ detection accuracy. The experimental results show that our method outperforms the state-of-the-art methods, including advanced small object detectors and aerial image detectors. Also, our multi-scale inference can increase the AP significantly while adding only a few processing pixels. If we downsample the input image to a lower resolution (e.g., 416 pixels with its short edge), a more efficient detector can be obtained. Some representative inference procedures and results are shown in Figure 12.

2) DOTA

The experimental results on the DOTA [7] dataset are displayed in Table 7. The initial input resolution of the image used in the method [16] is $600 \times 1000$, and the summation of global images and cropped chips are similar to their evaluation metric. We convert them to average the number of processing pixels per image to compare them fairly with the proposed method. EIP indicates evenly image partition (EIP), which can improve the utilization of image resolution directly. Our method achieves better performance than the popular detectors. Compared with ClusDet [16], which also uses an efficient multi-scale inference method in aerial images, our method can achieve better detection performance and further reduce the processing pixels. Although multi-scale can increase the number of samples effectively, it also introduces extreme scale noise. We use metric learning to distinguish extreme scale objects, which improves multi-scale training and inference. Figure 13 shows some representative inference procedures and results. The proposed algorithm has a significant efficiency improvement for some dense object areas such as parking lots and gathering areas.

3) UAVDT

The experimental results on the UAVDT [17] dataset are displayed in Table 8. All compared methods’ performance
is lower than VisDrone and Dota because of unbalanced data and poor label. The annotation of UA VDT is inaccurate, and several objects are missing. Furthermore, compared with the two other datasets, UAVDT has fewer small objects and classes, which leads to less promotion on our method. UAVDT’s objects mainly appear in the center of the image.
FIGURE 13. Qualitative detection results on DOTA dataset, each column shows the inference results of different scenes with an initial resolution of 416 × 416. The results shown in each row are the same as in Figure 12.

The EIP operation divides the objects into pieces such that the detector cannot estimate the scale of the objects correctly, thereby decreasing the performance dramatically.

On the contrary, our method is superior to EIP and ClusDet [16]. On the one hand, the image is cropped on the basis of clusters and contextual information, which is less likely to truncate numerous objects. Lastly, the detector can grasp the scale scaling region better, and multi-scale training is more efficient because of the use of extreme scale metric learning.

V. CONCLUSION
In this work, we analyze the problem of extreme scale in multi-scale training and inference of aerial image object detection and propose an extreme scale discrimination algorithm based on metric learning. Compared with the softmax classifier, pair-based metric learning can select extreme scale objects according to the unified distance, which improves the actual multi-scale detection performance. In addition, we use extremely small-scale regions for efficient multi-scale inference, which is superior to current popular methods. The experimental results show that the proposed algorithm can achieve the current state-of-the-art in terms of accuracy and the number of processing pixels on three popular aerial image datasets, namely, VisDrone, DOTA, and UAVDT.

In practical applications, the proposed algorithm can be combined with lightweight networks (e.g., MobileNet [51]) and pruning and quantization technologies (e.g., Nvidia Tensor RT) to achieve faster object detection in airborne aerial images.

ACKNOWLEDGMENT
The authors would like to express their sincere appreciation to the editor for his/her careful work in improving the presentation and quality of our article.

REFERENCES
[1] Y.-C. Du, M.-X. Zhang, H.-F. Ling, and Y.-J. Zheng, “Evolutionary planning of multi-UAV search for missing tourists,” IEEE Access, vol. 7, pp. 73480–73492, 2019.
[2] J. Sun, B. Li, Y. Jiang, and C.-Y. Wen, “A camera-based target detection and positioning UAV system for search and rescue (SAR) purposes,” Sensors, vol. 16, no. 11, p. 1778, Oct. 2016.
[3] V. P. Malagi, D. R. R. Babu, and K. Rangarajan, “Multi-object tracking in aerial image sequences using aerial tracking learning and detection algorithm,” Defence Sci. J., vol. 66, no. 2, pp. 122–129, 2016.
[4] Y. Jin, Z. Qian, and W. Yang, “UAV cluster-based video surveillance system optimization in heterogeneous communication of smart cities,” IEEE Access, vol. 8, pp. 55654–55664, 2020.

[5] H.-Y. Cheng, C.-C. Weng, and Y.-Y. Chen, “Vehicle detection in aerial surveillance using dynamic Bayesian networks,” IEEE Trans. Image Process., vol. 21, no. 4, pp. 2152–2159, Apr. 2012.

[6] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, “Vision meets drones: A challenge,” 2018, arXiv:1804.07437. [Online]. Available: http://arxiv.org/abs/1804.07437

[7] G.-S. Xia, X. Bai, D. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, “DOTA: A large-scale dataset for object detection in aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3974–3983.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, Feb. 2020.

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

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.

[11] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6577–6585.

[12] B. Singh, M. Najibi, and L. Davis, “Sniper: Efficient multi-scale training,” in Proc. 32nd Ann. Conf. Neural Inf. Process. Syst. (NIPS), 2018, pp. 9310–9320.

[13] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[14] M. Gao, R. Yu, A. Li, V. L. Morariu, and L. S. Davis, “Dynamic zoom-in network for fast object detection in large images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6926–6935.

[15] M. Najibi, B. Singh, and L. Davis, “AutoFocus: Efficient multi-scale inference,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9744–9754.

[16] 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. (ICCV), Nov. 2019, pp. 8310–8319.

[17] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, “The unmanned aerial vehicle benchmark: Object detection and tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 375–391.

[18] L. Mou and X. X. Zhu, “Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 11, pp. 6699–6711, Nov. 2018.

[19] H. He, Y. Lin, F. Chen, H.-M. Tai, and Z. Yin, “Inshore ship detection in remote sensing images via weighted pose voting,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 6, pp. 3091–3107, Jun. 2017.

[20] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, “HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 12, pp. 7147–7161, Dec. 2018.

[21] R. Touzi, J. Hurley, and P. W. Vachon, “Optimization of the degree of polarization for enhanced ship detection using polarimetric RADARSAT-2,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 10, pp. 5403–5424, Oct. 2015.

[22] K. Li, G. Cheng, S. Bu, and X. You, “Rotation-insensitive and context-augmented object detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[23] I. Sevo and A. Avramovic, “Convolutional neural network based automatic object detection on aerial images,” IEEE Geosci. Remote Sens. Lett., vol. 13, no. 5, pp. 740–744, May 2016.

[24] N. Audibert, B. Le Saux, and S. Lefèvre, “Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images,” Remote Sens., vol. 9, no. 4, p. 368, 2017.

[25] L. W. Sommer, T. Schuchert, and J. Beyerer, “Deep learning based multi-category object detection in aerial images,” in Proc. 27th Autom. Target Recognit., Anaheim, CA, USA, May 2017, Art. no. 1020209.

[26] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, “Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 8, pp. 3652–3664, Aug. 2017.
REN JIN received the Ph.D. degree from the School of Aerospace Engineering, Beijing Institute of Technology, Beijing, China, in 2020. He is currently a Postdoctoral Researcher with the School of Information and Electronics, Beijing Institute of Technology. His research interests include object detection, object tracking, remote sensing image processing, and machine learning.

JUNNING LV received the B.S. degree from the School of Aerospace Engineering, Beijing Institute of Technology, Beijing, China, in 2019, where he is currently pursuing the master’s degree. His main research interests include object detection, deep learning, and sensor fusion.

BIN LI received the Ph.D. degree from the School of Aerospace Engineering, Beijing Institute of Technology, Beijing, China, in 2019. He currently holds a postdoctoral position with the School of Aerospace Engineering, Beijing Institute of Technology. His main research interests include guidance systems, object detection, deep learning, and optimization method.

JIANCHUAN YE received the B.S. degree in flying vehicle design from the Beijing Institute of Technology, Beijing, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Aerospace Engineering. His main research interests include aircraft imaging system design, machine learning, and optimization method.

DEFU LIN received the Ph.D. degree from the School of Aerospace Engineering, Beijing Institute of Technology, Beijing, China, in 2004. He is currently a Professor of flight vehicle design with the Beijing Institute of Technology. He is also the Head of the Institute of UAV Autonomous Control and the Director of the Beijing Key Laboratory of UAV Autonomous Control with the Beijing Institute of Technology. His research interests include aircraft imaging system design, machine learning, and deep learning.