SDPDet: Learning Scale-Separated Dynamic Proposals for End-to-End Drone-View Detection

Nengzhong Yin, Chengxu Liu, *Graduate Student Member, IEEE*, Ruhao Tian, and Xueming Qian

***Abstract*—Detecting objects in large-scale drone-view images is notoriously challenging due to their uneven distribution and scale variation caused by photoing angles. Common approaches promote drone-view object detection by two-step detection (i.e., detecting sub-regions first) and multi-scale input. However, all these methods suffer from onerous computational costs since the high model complexity and input resolution. In this paper, we propose a novel one-step detector, called SDPDet, to enable effective object learning in drone-view images. In particular, a Scale-separated Activation Pyramid (SAP) serves to focus on the regions with objects aggregated at each scale, and a Scale-separated Learnable Proposals (SLP) mechanism learns proposal boxes and corresponding features on these regions. By such design, the quantity of learnable proposals allows dynamic adjustment at each scale separately, which facilitates the objects learning of various distributions and scales with less computational costs. Experiments demonstrate SDPDet can significantly outperform the state-of-the-art one-step detectors on three widely-used benchmarks. On the most challenging VisDrone dataset, SDPDet with ResNet50 gains 5.4% AP and 6.9% AP_s improvements while running 1.9× faster than previous models.**

***Index Terms*—Drone-view image, activation pyramid, scale-separated learnable proposals, object detection.**

I. Introduction

DRONE-VIEW image detection aims at classifying and locating objects in a large field of view captured by drones or surveillance cameras. It is a fundamental problem in computer vision and can be applied to numerous applications, including agriculture [1], security surveillance [2], and rescue search [3]. These applications require robust and efficient detectors. Benefiting from the development of various generic object detection datasets (e.g., MS COCO [4], Pascal VOC [5], Objects365 [6]), object detection has developed at a rapid pace and has a wide range of applications in various fields [7], [8], [9]. Many state-of-the-art detectors based on CNNs have shown excellent performance. Such as the Faster-RCNN [10], and YOLO series [11]. However, these detectors are mostly designed for generic object detection. In drone-view scenarios such as VisDrone-DET [12] and UAVDT [13], there are more objects in images, with more extreme scale variation and more uneven distribution, thus limiting the performance of these detectors. From the perspective of practical scenarios, drone-view images suffer from scale and distribution problems: 1). The objects in the drone view image are unevenly distributed and with fewer areas than the background. There are some areas in the images with a high density of objects, while others are mostly without objects. 2). In drone-view images, the scale varies greatly even for the same category of objects from different shooting angles, and the scale disparity is even more pronounced for different categories of objects, such as trucks and people. Therefore, it is necessary for the detector to avoid invalid computations on regions without objects and have the ability to handle multiple scale variance.

To solve these challenges, recent years have witnessed an increasing number of drone-view object detection approaches, which can be categorized into two paradigms: two-step scheme and one-step scheme. The former attempts a two-step scheme from coarse to fine detection [14], [15], [16], [17], [18]. As shown in Fig. 1(a), they first use a coarse detector to obtain sub-regions that object aggregation, then use a fine detector to detect them from these regions. One of the classic works is GLSAN which introduces super-resolution networks to scaling sub-regions to the proper size for better fine detection [16]. However, this paradigm is time-consuming and complex, especially when it encounters scenarios with discrete object distributions, the paradigm will dramatically increase the computational costs and reduce efficiency. The latter detects objects directly through a one-step scheme mainly by using high-resolution features or performing feature fusion [19], [20], [21]. To avoid invalid computational costs for regions without objects, the latest QueryDet proposes a query mechanism to locate small objects on high-resolution features [21]. Nonetheless, the NMS post-processing necessary for such box/point anchor-based detectors is still inefficient in handling objects that are dense and multi-scale objects.

Inspired by the recent progress of Transformer in computer vision [22], significant progress has been made in generic object

Manuscript received 14 April 2023; revised 8 January 2024; accepted 24 February 2024. Date of publication 5 March 2024; date of current version 24 April 2024. This work was supported in part by NSFC under Grant 62272380 and Grant 62103317, in part by the Fundamental Research Funds for the Central Universities, China under Grant xzy022023051, and in part by the Innovative Leading Talents Scholarship of Xi'an Jiaotong University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. F. Sohel. *(Nengzhong Yin and Chengxu Liu contributed equally to this work.) (Corresponding author: Xueming Qian.)*

Nengzhong Yin and Chengxu Liu are with the School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: yinnz@foxmail.com; liuchx97@gmail.com).

Ruhao Tian is with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: ruhaot2020@stu.xjtu.edu.cn).

Xueming Qian is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with SMILES LAB, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TMM.2024.3371892

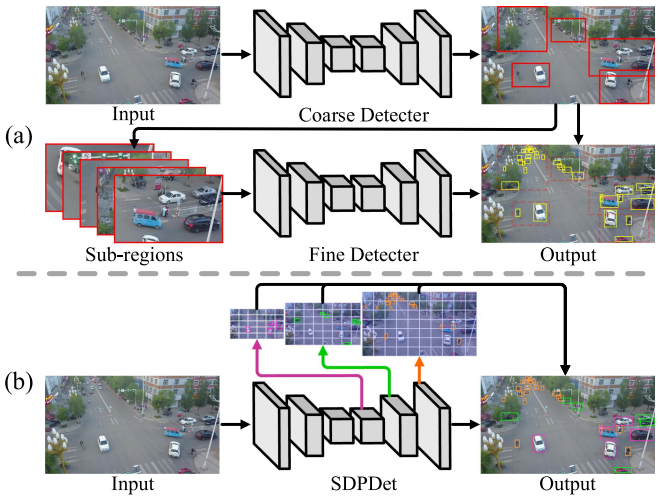

Fig. 1. Comparison of different pipelines. (a) Two-step detection usually first detects the sub-regions containing small objects and then detects the category and the location of each object in these sub-regions. (b) Our proposed SDPDet utilizes the feature pyramids to construct dynamic learnable proposals at each scale separately for object learning.

detection [23], [24], [25]. For example, DETR [23] constructs a sparse set of object queries in Transformer to reason about the relations of the objects and the global image context for obtaining the final prediction set. To further avoid dense feature interactions, Sparse R-CNN [24] only uses a pre-defined set of sparse learnable proposals to learn object positions and categories directly. The mechanism of learnable proposals learns the distributional of objects in the dataset and alleviates the problem of uneven distribution in the drone-view image. However, the way of learning all proposals in a uniform set degrades the ability when handling uneven distributions and scale variations of objects, leading to sub-optimal performance. Therefore, exploring proper ways of utilizing the learnable proposal mechanisms in drone-view object detection remains a big challenge.

Addressing both challenges above, we propose a novel **S**cale-separated **D**ynamic **P**roposals for End-to-End Drone-View Object **Det**ection, called **SDPDet**, which can automatically adjust the number of proposals according to the object distribution and separate the proposals to learn the scale information specifically. As shown in Fig. 1(b), SDPDet utilizes the feature pyramid to construct scale-separated dynamic learnable proposals, i.e. , the quantity of learnable proposals allows dynamic adjustment at each scale separately, to learn objects with uneven distribution and scale variation. To achieve this purpose, as shown in Fig. 2, we first construct a scale-separated activation pyramid (SAP) to guide the proposals learning in the regions with objects aggregated at each scale, avoiding the invalid costs due to uneven distribution. Then we propose a scale-separated learnable proposals (SLP) mechanism to separate the proposals into scaled groups, so each group learns only a certain scale of the objects. This separation mechanism makes different levels of features learn the scale feature exclusively, which can avoid sub-optimal performance caused by scale variations.

Compared with the latest one-step detection methods, the proposed SDPDet not only significantly avoids the computational costs of no-object regions, but also optimizes the learning of objects at different distributions and scales.

Our contributions are summarized as follows:

- We propose a novel one-step detector, called SDPDet, which is the first work to introduce the learnable proposals mechanism into drone-view detection. It can significantly alleviate the problem of uneven distribution and multi-scale in drone-view images with higher efficiency.
- We propose a scale-separated learnable proposal (SLP) mechanism, which can dynamically adjust the learnable proposals at each scale to optimize the object learning of different scales.
- We propose a scale-separated activation pyramid (SAP) to enable the model to focus on regions of object aggregation in images with uneven distribution, significantly reducing the computational cost of no-object regions.
- Extensive experiments demonstrate the superiority of the SDPDet over state-of-the-art one-step detectors on three widely-used benchmarks. On the VisDrone dataset, SDPDet with ResNet50 gains **5.4%** AP improvements and runs **1.9**× faster.

II. Related Works

In this section, we first briefly introduce the object detection methods for general scenarios in Section II-A, and then describe the related work of drone-view object detection in detail in Section II-B.

A. Generic Object Detection

According to the different post-processing, generic object detection can be categorized into two paradigms. NMS-based algorithms [10], [11], [26], [27] and NMS-free algorithms [23], [24], [25], [28].

NMS-based methods: The development of these approaches can be divided into two main groups: two-stage methods, one-stage methods. The two-stage methods [10], [29] first generate a region of interest (RoI) that may contain objects, and then further classify and regress these regions to obtain the final detection results, such as Faster R-CNN [10], Cascade R-CNN [29]. To accelerate the detection speed, the one-stage methods [11], [30], [31], [32] predict the bounding box directly by computing the extracted features. Both anchor-based (e.g., SSD [30], YOLO [11]) and anchor-free (e.g., CenterNet [32], FCOS [33]) methods are included. However, all of the above methods regress bounding boxes by creating dense candidates at fixed locations and require NMS post-processing to remove redundant predictions during inference. However, when objects are extremely dense, the predictions of adjacent objects overlap too much so that the correct result may be removed by NMS.

NMS-free methods. NMS-free [23], [24], [25], [28] detectors reformulate object detection as a set prediction problem. They designed a small number of learnable object queries to model the relationship between objects and the global image and showed impressive performance. Typically, the DETR series (e.g., DETR [23], Anchor DETR [28], Deformable DETR [25]) delicately introduces the Transformer to object detection and

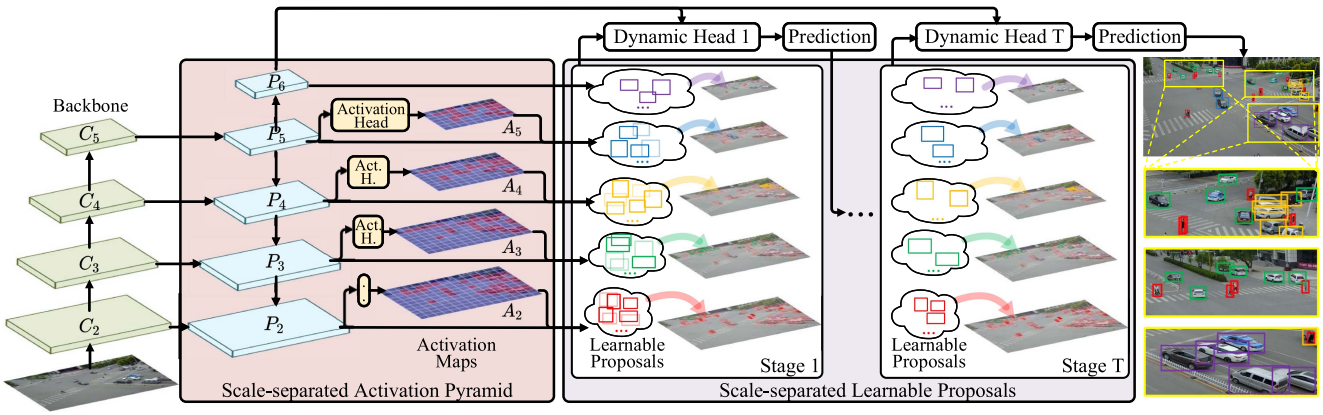

Fig. 2. Overview of SDPDet. The input image generates multi-scale features by the backbone. Scale-separated activation pyramid (SAP) is used to focus on the regions with objects aggregated in a scale-separated manner. Scale-separated learnable proposals (SLP) mechanism learns the regression and classification on different layers separately within the region of activation in multiple stages. Dynamic head is used to refine the prediction among the stages and output the final detection results without any post-processing.

allows the network to focus on object regions by improving the design of the query. Sparse R-CNN [24] proposes a learnable proposals mechanism, which learns the location of objects to extract RoI. Methodologically, compared with creating dense candidates in all regions, using such a learnable proposals mechanism reduces more computational costs, especially for objects with discrete distributions.

B. Drone-View Object Detection

Drone-view object detection aims to detect objects with a more discrete distribution and scale variation in a high image resolution. Typical drone-view object detectors are mainly divided into two paradigms, two-step and one-step detection.

Two-step detection: To avoid the invalid computational costs introduced by the discrete distribution of objects and the correct results removed by NMS for dense aggregation, this paradigm [14], [15], [16], [17] first searches a region where objects aggregated by a coarse detector, and then detects them with a fine detector. Most of these approaches either optimize the object learning by increasing the quantity of samples through data augmentation [34], [35] or utilize generative adversarial network (GAN) [36], [37], [38] to enhance the representation of the objects. Some typical approaches (e.g., ClusDet [14], DMNet [15], GLSAN [16]) introduce clustering algorithm following the coarse detector to locate the object aggregated regions. Although achieving high accuracy, they generate huge computational costs for searching many sub-regions due to extreme distributions in drone-view images.

One-step detection: To further improve efficiency, one-step detectors directly detect the objects in the drone-view images. The existing approaches mainly facilitate object learning by increasing the resolution [19], [20], [21], [39], contextual learning [40], [41], [42], and multi-scale learning [20], [21], [43], [44], [45]. Among these approaches increasing the image resolution is the most effective, but it significantly increases the computational costs. The recent QueryDet [21] reduces computational costs while using higher resolution features by introducing the query mechanism. Nonetheless, this box/point anchor-based detector is still challenging to handle objects that are dense and contain more than one scale at the same time.

In this paper, we introduce the learnable proposals mechanism into the one-step detection paradigm to solve the above problem in a more efficient way.

III. Preliminary

Previous methods that regress bounding boxes based on the pre-defined box/point anchors [10], [30], [43]. These methods require dense candidates and NMS post-processing to remove redundant predictions during inference. However, in handling dense and multi-scale object scenarios, NMS usually filters out some correct prediction results that overlap too much with adjacent predictions.

To solve these issues, the recent learnable proposals-based detectors [24], [46] model regression and classification in detection as a set prediction problem, which produces an optimal bipartite matching between ground truth (GT) and predictions without NMS post-processing. Specifically, these methods are similar to Faster R-CNN [10], which first extract features by the backbone, then obtain proposals and further refine them to get bounding boxes and classification. Rather than generate proposals by the Region Proposals Network (RPN), they propose the learnable proposals mechanism. It regresses and classifies objects by a set of learnable proposal boxes and learnable features. Each learnable proposal box contains four parameters (i.e. , the coordinates of the box center, and the width and height of the box) to obtain the Region of Interest(RoI) in the proposal box area, and each learnable proposal feature contains a one-dimensional feature that interacts with the RoI to assist the regression and classification. The proposal boxes and features can be progressively refined to output prediction results in multiple stages by dynamic heads during inference and also can be updated by back-propagation during training.

Although these methods have promising performance in general object detection, they fail to handle the severe scale variation in drone-view object detection. This is because this binary matching-based approach is sensitive to the dataset. The high

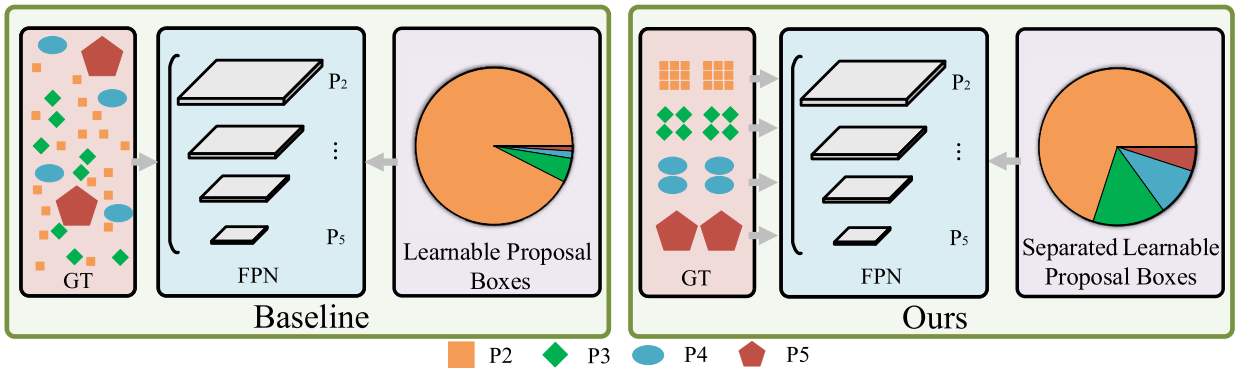

Fig. 3. Difference between SLP and vanilla learnable proposals. SLP separates proposals into scale groups, the number of each group is determined by the distribution of the dataset, and all GTs are involved in the training of specific levels according to their scale. In this way, the RoI from each level can be sufficiently extracted during training for better exploiting the advantages of FPN.

proportion of small-scale objects in the drone-view images will cause the learnable proposals to be biased toward small-scale features and neglect other scales.

IV. Methods

In this section, we first introduce the proposed SDPDet in Section IV-A, and then we describe in detail the two components of SDPDet, scale-separated learnable proposals (SLP) in Section IV-B and scale-separated activation pyramid (SAP) in Section IV-C, respectively. Finally, the training details are introduced in Section IV-D.

A. SDPDet

The overview of our proposed SDPDet is shown in Fig. 2. Firstly, we feed the input image into the backbone to obtain the multi-scale features (i.e. , $C_2 \sim C_5$). Then, the multi-scale features are fed into the SAP to obtain the pyramid features (i.e. , $P_2 \sim P_6$) and activation maps (i.e. , $A_2 \sim A_5$). Next, the SLP mechanism enables each proposal to learn within the activation regions at the corresponding scale and extract RoI on different scales. Finally, we sent the RoI and its corresponding proposal feature to the dynamic head same as existing work [24], [46] for obtaining prediction results at multiple stages. The proposed SLP and SAP are described in detail in Sections IV-B and IV-C, respectively.

The **key idea** of SDPDet is the scale-separated dynamic proposals mechanism, which is an effective combination of SAP and SLP. Specifically, we determine whether the learnable proposal box from level l contains the points in the activation map A_l from SAP. If so, we assume that the proposal contains objects and preserve it, otherwise discard it. The preserved proposals are refined through a total of T stages to output prediction results. With this scale-separated dynamic proposals mechanism, we can dynamically preserve proposals that contain objects at each scale separately and avoid invalid ones. The **advantages** of SDPDet are to optimize the object learning at each scale while reducing the invalid computational costs and solving the problem of uneven distribution and scale variation in drone-view images.

B. Scale-Separated Learnable Proposals

The Vanilla learnable proposal mechanism [24] treats all GTs as a uniform set to learn during training, and each proposal will be assigned by its scale to different levels for extracting the region of interests (RoI). As shown in the left part of Fig. 3, in drone-view images dominated by small objects, almost all the proposals are assigned to the feature map with higher resolution P_2, and almost no proposals are available in the remaining levels. Besides, in different images, the same proposals will learn different scales of objects which leads to sub-optimal solutions.

Therefore, we propose the scale-separated learnable proposals (SLP) mechanism. In particular, inspired by the FPN's idea of divide-and-conquer, we consider each single learnable proposal as an 'individual' and all proposals as a collective 'whole'. Each 'individual' focuses on only one scale of feature learning, which constitutes a more complete and effective 'whole' for scale learning.

As shown in Fig. 3, we decouple the learnable proposals at each scale and train them separately with separated GT on different layers of the FPN [43]. Specifically, during training, for a bounding box with width w and height h, we assign it to the level P_l of the feature pyramid by:

$$l = \begin{cases} 2, & \sqrt{wh} \le \frac{1}{4}s \\ \lfloor l_0 + \log_2 \frac{\sqrt{wh}}{s} \rfloor, & \frac{1}{4}s < \sqrt{wh} \le 4s\,, \\ 6, & 4\,s < \sqrt{wh} \end{cases} \tag{1}$$

where l refers to the pyramid level. $\lfloor \cdot \rfloor$ denotes the floor function. We empirically follow existing work that adapts the assignment strategy of FPN-based detectors [10], [43] setting s to 224. l_0 is the target level on which the RoI with $w \times h = 224^2$ should be mapped into, and l_0 is set to 4 for the ResNet-based object detection framework.

Based on the assignment results of the proposal boxes in the FPN during training, each different proposal only focuses on the prediction at the corresponding scale during inference. By such design, we make the proposals robust to scale variation and focus on learning different scale objects in a separate way.

C. Scale-Separated Activation Pyramid

Due to the uneven distribution of objects, detecting in drone-view images usually produces onerous and invalid computational costs on the regions without objects, especially for shallow features with higher resolution of FPN. Inspired by this, we propose the scale-separated activation pyramid (SAP), which activates regions of object aggregation in the feature pyramid at each scale separately by designing a simple activation head.

Specifically, for an image with input size $h \times w$, we use $\mathcal{P} = \{P_l \in \mathbb{R}^{h_l \times w_l \times c_l}\}$ to denote the feature maps output from SAP, where l refers to the pyramid level, c_l is the feature channel of level l, and (h_l, w_l) refers to $\left(\frac{h}{2^l}, \frac{w}{2^l}\right)$. We input the features P_l at the l level of SAP for generating the activation maps $A_l \in \mathbb{R}^{h_l \times w_l \times 1}$ at each level separately. The detailed process can be divided into the following two steps. 1) Feeding P_l into the activation head consisting of several convolutional layers to obtain a single-channel heatmap, which represents the confidence of the aggregated region presence. 2) Generating activation map by setting the regions in the heatmap with values larger than the threshold T_a to 1 and the rest to 0.

During training, based on the assignment strategy in (1), we enable the activation head at each level to only learn the regions with objects aggregated at the corresponding scale. In this simple but effective way, SAP focuses on the regions where object aggregation in drone-view images at each scale better, while consuming very little costs.

Unlike QueryDet [21], which introduces an additional query head branch after the output of the feature pyramid to detect the small objects coarsely. It uses sparse convolution [47] to reduce the computational cost by computing only the areas of small objects in high-resolution features. The learnable proposals mechanism does not use convolutional operations on the full image, which means there is no need to use sparse convolution to reduce computational costs. Therefore, we concentrate more on how to obtain robust aggregation regions in parallel with the output feature pyramids. Specifically, the activation head in our SAP is to detect the region of object aggregation separately at each scale, so as to avoid invalid calculations without object regions.

D. Training

We divide the loss into two parts, one part used to guide the activation maps generation in the SAP, and another part used to supervise the classification and regression of SLP.

1) SAP: To enable the generated activation maps to focus on regions with objects aggregated, we construct pseudo-labels based on GT and train the activation head. Specifically, we denote each object as $b = (x_c, y_c, w, h)$, where (x_c, y_c) as the center, and w and h are the width and height, respectively. We construct the circle with (x_c, y_c) as the center and $r = \sqrt{(w/2)^2 + (h/2)^2}$ as the radius. The GT of the activation map A_{GT} can be formulated as:

$$A_{GT} = \bigcup_{x,y} A(x, y)$$

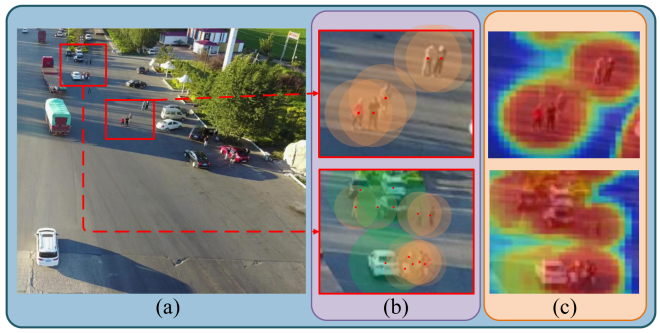

Fig. 4. We use circles covering the objects as labels. When multiple objects gather, multiple circles overlap with each other to obtain higher confidence, thus obtaining the aggregated region. (a) Original image. (b) Ground-truth of SAP. (c) Visualization activation map obtained by SAP.

$$A(x, y) = \begin{cases} 1 & if\ D(x, y) < r_a \\ 0 & if\ D(x, y) \geq r_a \end{cases}, \tag{2}$$

where $D(x, y) = \sqrt{(x - x_c)^2 + (y - y_c)^2}$ refers to the Euclidean distance between the point (x, y) in A_{GT} and the center of the circle (x_c, y_c), the r_a refers to a multiple of r and uses to measure the distance between aggregated objects. The loss of activation maps generation $\mathcal{L}_a$ can be defined as:

$$\mathcal{L}_a = FL(A, A_{GT}), \tag{3}$$

where A is the activation map output from the SAP. $FL(\cdot)$ denotes the FocalLoss [31]. As shown in Fig. 4, the network is trained to obtain a circle region centered on the objects, and the aggregated regions obtain higher confidence scores due to the circles of objects overlapping each other. It is worth noting that this design of generating A_{GT} can be applied at different scales to allow the training of activation heads in the SAP.

2) SLP: We use the same loss design as the set prediction paradigm approach [23], [24], [46] for SLP. Specifically, this loss function supervises the learnable proposals by generating an optimal bipartite matching between the prediction and GT. The loss $\mathcal{L}_p$, i.e. , the matching cost, can be defined as:

$$\mathcal{L}_p = \lambda_{cls} \cdot \mathcal{L}_{cls} + \lambda_{L1} \cdot \mathcal{L}_{L1} + \lambda_{giou} \cdot \mathcal{L}_{giou}, \tag{4}$$

where $\mathcal{L}_{cls}$ denotes the FocalLoss [31] between predicted classification and GT label. $\mathcal{L}_{L1}$ and $\mathcal{L}_{giou}$ represent the $L1$ loss and generalized IoU loss [48] between predicted bounding box and GT box, respectively. λ_{cls}, λ_{L1}, and λ_{giou} are the coefficients of three part. To optimize the feature learning in different layers better, we calculate the loss function in each layer with only matched pair that satisfies the scale in (1). i.e. , The proposals on P_2 will only match the smallest object and vice versa.

Finally, the overall loss $\mathcal{L}_{total}$ can be represented as:

$$\mathcal{L}_{total} = \mathcal{L}_a + \mathcal{L}_p, \tag{5}$$

where $\mathcal{L}_a$ is the activation maps loss at all scales, and $\mathcal{L}_p$ is the classification and regression loss at all scales.

V. Experiments

A. Implementation Details

To trade off the complexity and accuracy of our model, we construct the activation head using five convolutional layers of size 3×3. Considering the runtime and the FLOPs, we filter out the proposals at P_2 and set the threshold T_a to 0.3. We set r_a to 2 times r to better obtain the regions with objects aggregated. Besides, to obtain the optimal proposals through the multi-stage refinement mechanism in SLP, we follow previous works [24], [46] to set T to 6. We use the AdamW optimizer with weight decay 0.0001, and use the batch size of 1 for 50 epochs. The initial learning rate is set as 2.5×10^{-5} and then decreases by a factor of 10 for epochs 30 and 40. We follow previous work [24], [46] to perform data augmentation by random horizontal, and scale jitter of resizing the input images. We follow previous works [23], [24], [25] to set the coefficients λ_{cls}, λ_{L1}, and λ_{giou} as 2, 5, and 2, respectively. Our proposed SDPDet is based on the Detectron2 toolkit and PyTorch, and all models are conducted on an NVIDIA RTX 2080Ti GPU.

B. Datasets and Metric

1) Datasets: To demonstrate the superiority of SDPDet, we validate performance on three widely-used drone-view image and remote sensing detection benchmarks, VisDrone [12], UAVDT [13], and DOTA [55].

VisDrone is acquired by drones at different viewpoints and altitudes, including 6,471 training images, 548 validation images, and 3,190 test images. The dataset has 10 categories, and its resolution is about 2000×1500. The object size can be divided into small ($area < 32^2$), medium ($32^2 \leq area < 96^2$), and large ($area \geq 96^2$). And the proportions in the dataset are [60.5, 34.0, 5.5]. For fair comparisons, we follow previous works [14], [15], [16] to evaluate SDPDet on the validation set.

UAVDT is a popular drone-view image dataset, which contains 23,258 images for training and 15,069 images for testing. It mainly contains three categories of objects, and the average resolution is 1080×540.

DOTA is the constructed dataset by selecting from the remote sensing benchmark DOTA [55]. For fair comparisons, we follow previous works [14], [18] to select 920 images for training and 285 images for testing, which included movable objects, such as airplanes, helicopters, *etc*.

2) Evaluation Metric: We follow existing works [14], [15], [16], [20], [21] and use the widely-used evaluation protocol of the MS COCO dataset. It involves six metrics AP, $AP_{0.5}$, $AP_{0.75}$, AP_s, AP_m, AP_l, where AP denotes the average precision for ten IoU thresholds whose range is from 0.5 to 0.95, while $AP_{0.5}$ and $AP_{0.75}$ are 0.5 and 0.75, respectively. AP_s, AP_m, and AP_l are the AP for the objects with $area < 32^2$, $32^2 \leq area < 96^2$, and $area \geq 96^2$, respectively.

C. Comparisons With State-of-the-Art Methods

We compare SDPDet with 13 state-of-the-art methods and categorize these methods according to the number of detectors used: two-step detection methods [14], [15], [16], [17], [18], [49], [50], [51], [52], [53], [54] and one-step detection methods [20], [21]. For fair comparisons, we obtain the performance from their original paper or reproduce results by authors' officially released models. We use the runtime of an image, denoted as s/img, to verify the model complexity.

We compare SDPDet with other SOTA methods on the most widely-used VisDrone dataset [12]. As shown in Table I, the upper and lower parts of the table show the results of the two-step detection methods and the one-step detection methods, respectively. Among them, since the strategy of detecting regions and slicing original images for training and testing, the two-step detection [14], [15], [16], [17], [18], [49], [50], [54] achieves higher accuracy, they have a more complex structure and longer runtime than one-step detection, especially when it encounters scenarios with the discrete distribution. Recent years have witnessed an increasing number of one-step detection methods [20], [21], which detect information-limited objects directly through one detector. Such as Focus&Detect [54], although it achieves the highest performance, the heavy structure and post-processing make its poor runtime. HRDNet [20] and QueryDet [21] attempt to optimize the feature extraction and runtime when high-resolution images are input. Nonetheless, these methods based on box/point anchors are still challenging to handle objects with uneven distribution and scale variation. Different from them, SDPDet tries to solve these problems by linking the regions of object aggregation and the learnable proposals mechanisms together in a scale-separated manner. Due to such merits, using the same backbone, SDPDet achieves a result of 33.7 AP and significantly outperforms QueryDet [21] by **5.4** AP on the VisDrone [12] and runs **1.9**$\times$ faster (0.196 s/img $vs.$ 0.364 s/img), and also has better performance than HRDNet [20] using a larger backbone (33.7 AP using ResNet50 $vs.$ 31.4 AP using ResNet18+101). This large margin demonstrates the power of SDPDet in uneven distribution and scale variation.

To further verify the generalization ability of SDPDet, we evaluate SDPDet on other two popular drone-view and remote sensing image detection benchmarks, UAVDT [13] and DOTA [55]. For fair comparisons, we follow previous works [14], [15], [16], [17], [18], [50], [55] to use the same training and testing strategy on the dataset. As shown in Table II, due to the well-designed dynamic adjustment mechanism of the learnable proposals, SDPDet achieves better results in all datasets. It is worth noting that since the images in DOTA are too large, two-step detection can detect by slicing the image from the detected sub-regions, but the images for one-step detection are too large to fit, so we use the official toolkit [55] to slice the images. The results verify that our SDPDet has strong generalization capabilities under different scenarios.

D. Ablation Study

To analyze how SAP and SLP influence the accuracy and complexity of SDPDet, we conduct ablation studies on the VisDrone validation set. We introduce s/img and FLOPs to verify the runtime and complexity. The results are shown in Table III. After introducing the SAP, we get a better performance in both accuracy (+0.8 AP) and complexity ($-$0.017 s/img and $-$34.6 G

TABLE I
RESULTS IN VISDRONE VALIDATION SET

Paradigm	Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_l	s/img
Two-step Detection	ClusDet [14]	ResNeXt101	28.4	53.2	26.4	19.1	40.8	54.4	0.773 (GTX 1080Ti)
	DMNet [15]	ResNeXt101	29.4	49.3	30.6	21.6	41.0	56.9	0.610 (GTX 1080Ti)
	CRENet [49]	Hourglass104	33.7	54.3	33.5	25.6	45.3	**58.7**	0.901 (RTX 2080Ti)
	CDMDet [17]	ResNeXt101	31.9	52.9	33.2	23.8	43.4	45.1	-
	GLSRN [16]	ResNet101	30.7	55.6	22.9	-	-	-	0.760 (TiTAN Xp)
	UCGNet [50]	DarkNet53	32.8	53.1	33.9	-	-	-	-
	ADaZoom [18]	ResNeXt101	37.6	**66.3**	39.5	-	-	-	-
	UFPMP-Det [51]	ResNeXt101	40.1	66.8	41.3	-	-	-	-
	PRDet [52]	ResNeXt101	32.0	53.9	33.2	25.6	40.8	52.9	0.195 (RTX 3090)
	CZ Det [53]	ResNet50	33.2	58.3	33.2	26.1	42.6	43.4	**0.118** (A100)
	Focus&Detect[54]	ResNeXt101	**42.1**	66.1	44.6	**32**	**47.9**	54.5	1.362 (RTX 2080Ti)
One-step Detection	QueryDet [21]	ResNet50	28.3	48.1	28.8	19.8	35.9	40.3	0.364 (RTX 2080Ti)
	HRDNet [20]	ResNet18+101	31.4	53.3	31.6	-	-	-	0.357 (RTX 2080Ti)
	SDPDet(Ours)	ResNet50	33.7	56.6	34.3	26.7	42.9	**45.7**	**0.196** (RTX 2080Ti)
	SDPDet(Ours)	ResNeXt101	**34.2**	**57.8**	**34.9**	**27.5**	**43.2**	41.9	0.346 (RTX 2080Ti)

Bold font indicates the best result of one column.

TABLE II
RESULTS IN UAVDT AND DOTA DATASET

Dataset	Paradigm	Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_l
UAVDT	Two-step Detection	ClusDet [14]	ResNet50	13.7	26.5	12.5	9.1	25.1	13.2
		DMNet [15]	ResNet50	14.7	24.6	16.3	9.3	26.2	**35.2**
		CDMDet [17]	ResNet50	20.7	35.5	22.4	**13.9**	**33.5**	19.8
		GLSRN [16]	ResNet50	19	30.5	21.7	-	-	-
		UCGNet [50]	DarkNet53	19.1	36.7	18	11.1	31	36
		ADaZoom [18]	ResNet50	19.6	33.6	21.3	14.4	28.6	31.2
		CZ Det [53]	ResNet50	19.8	34.1	21.3	-	-	-
		UFPMP-Det [51]	ResNet50	**24.6**	**38.7**	**28.0**	-	-	-
	One-step Detection	QueryDet [21]	ResNet50	14.3	27.2	16.6	11.1	24.2	14.7
		SDPDet(Ours)	ResNet50	**20.0**	**32.0**	**23.1**	**13.3**	**33.0**	**21.1**
DOTA	Two-step Detection	ClusDet* [14]	ResNet50	32.2	47.6	39.2	16.6	32	**50**
		AdaZoom* [18]	ResNet50	**36.0**	**62.7**	**37.0**	-	-	-
		CZ Det* [53]	ResNet50	34.6	56.9	36.2	**18.2**	**37.8**	43.8
	One-step Detection	QueryDet [21]	ResNet50	33.9	58.1	36.3	16.3	40.3	39.4
		SDPDet(Ours)	ResNet50	**40.9**	**62.3**	**47.8**	**22.4**	**47.8**	**53.5**

* indicates using the uncropped original image as input.
SAP stands for scale-separated activation pyramid, SLP stands for scale-separated learnable proposals.

TABLE III
ABLATION STUDIES ON VISDRONE VALIDATION SET

SAP	SLP	AP	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_l	s/img	FLOPs
		29.9	50.9	30.5	23.4	38.4	**45.9**	0.177	155.8G
✓		30.7	52.0	31.5	24.2	40.0	37.4	**0.160**	**121.2G**
	✓	33.3	56.1	34.3	26.0	42.9	44.9	0.215	175.1G
✓	✓	**33.7**	**56.6**	**34.3**	**26.7**	**42.9**	45.7	0.196	139.7G

SAP stands for scale-separated activation pyramid, SLP stands for scale-separated learnable proposals.

FLOPs). It demonstrates that SAP can improve both accuracy and efficiency by enabling the detector to focus on the aggregated region and filter out invalid proposals. In addition, the utilization of aggregated region information during training the SAP has led to significant improvements in the detection performance of smaller objects. This was achieved by augmenting the detector's heightened attention toward such regions where smaller objects aggregated. The introduction of the SLP mechanism can largely increase accuracy (+3.4 AP), especially in small objects (+2.6 AP_s and +4.5 AP_m). It proves that by scale-separating learnable proposals, the learnable proposals can learn scale information to handle object uneven distribution and scale problems more effectively. However, the mechanism of SLP increases the runtime and computational costs. After combining SAP and SLP in SDPDet at the same time, we get a further performance with a gain of 3.8 AP and 16.1 G fewer FLOPs, which proves the superiority of our proposed SAP and SLP.

E. Visualization

To verify the effectiveness of the activation map generated by the activation head. As shown in Fig. 5, we visualize it and its corresponding detection results. Regardless of whether the objects are dense or discrete, the activation map focuses on the regions with objects aggregated, which indicates the superiority of the activation map.

VI. DISCUSSIONS

In this section, we analyze each factor in SAP and SLP to prove their validity as much as possible. All experiments are conducted on the VisDrone [12] validation set.

A. Influence of the Activation Map Threshold T_a

As described in Section IV-C. The activation map is used to search the regions with objects aggregated and filter proposals

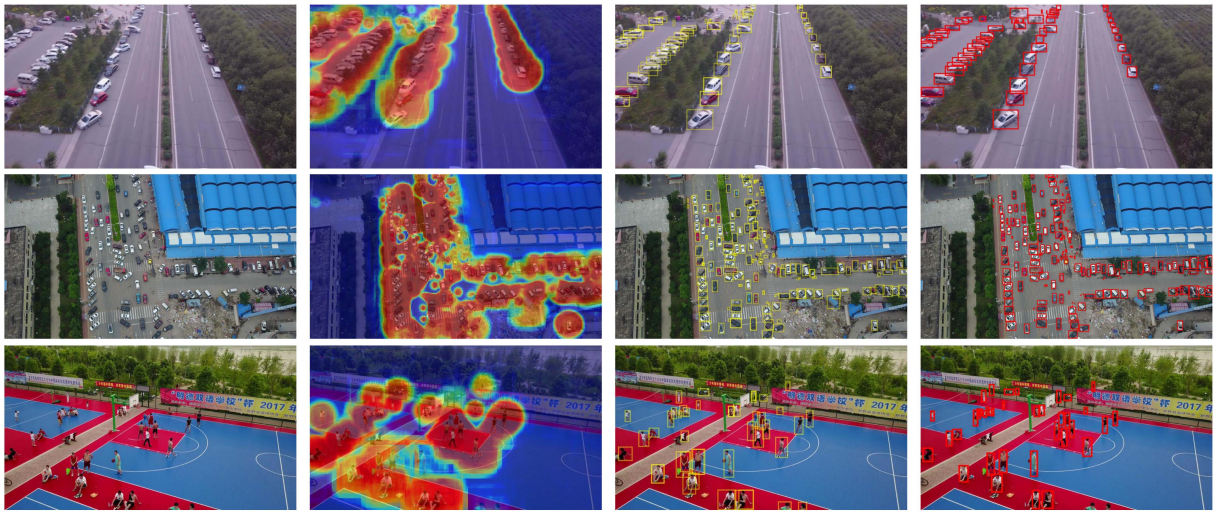

Fig. 5. Visualization of the activation map and detection results on VisDrone validation set, We remove the category labels in the bounding box for a better visual experience.

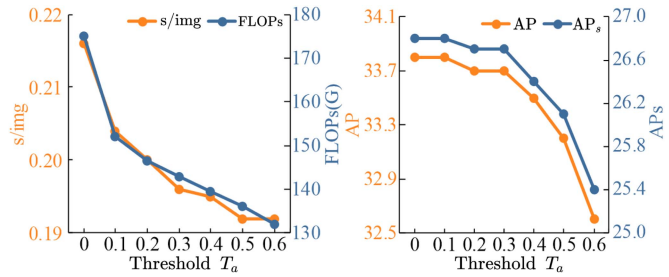

Fig. 6. Sensitivity of activation map thresholds for model accuracy (left) and complexity (right).

in regions smaller than the threshold T_a to reduce computational costs. Therefore, we explore the sensitivity of the T_a in the activation map. As shown in Fig. 6, the effect of threshold T_a on runtime and computation costs is almost linear, while the effect on detection performance has little change up to 0.3 and starts to drop sharply when it exceeds 0.3. Setting T_a as 0.3 can filter out invalid proposals and reduce the complexity almost without reducing accuracy. Yet further increasing T_a would filter out the correct proposals and reduce the performance. We set the T_a to 0.3 after a trade-off.

B. Influence of Activation Map on Different Layers

To explore the influence of activation maps on complexity, we construct comparisons by adding activation maps on different layers of SAP separately. As shown in Table IV, compared with no activation map, using activation maps at the higher resolution P_2 can reduce much computational costs (-35.4 G FLOPs) without reducing accuracy. Since the proposals are computed in parallel, the reduction in runtime is not significant (-0.020 s/img). It is worth noting that the performance of AP$_l$ decreases by 3.3 when using the activation map at P_3, which is due to the fact that the object distribution is more discrete at this scale and the activation map causes the proposals lost to some extent. Since objects are much rarer in other scales, they have little impact on performance. After a trade-off between accuracy and complexity, we use the activation map at P_2 as the final model.

TABLE IV
COMPARISON OF PERFORMANCE AND COMPLEXITY USING ACTIVATION MAP IN DIFFERENT LAYERS

Layer	AP	AP$_s$	AP$_m$	AP$_l$	s/img	FLOPs
-	33.8	26.8	42.9	45.8	0.216	175.1G
P_2	33.7	26.7	42.9	45.7	0.196	139.7G
P_3	33.6	26.8	42.7	42.4	0.212	169.6G
P_4	33.7	26.8	42.9	45.2	0.215	173.9G
P_5	33.8	26.8	42.9	45.4	0.215	173.7G

TABLE V
COMPARISON OF DIFFERENT METHODS OF OBTAINING ACTIVATION MAP

Method	AP	AP$_s$	AP$_m$	AP$_l$	s/img	FLOPs
QueryDet	33.0	25.8	42.3	45.5	0.181	125.4G
Ours(r_a 2)	33.7	26.7	42.9	45.7	0.196	139.7G
Ours(r_a 4)	33.7	26.7	43.0	45.9	0.207	159.4G

C. Different Methods of Obtaining Activation Map

We apply our method and QueryDet to SAP for comparison, respectively. We use r_a to measure the distance between aggregation objects when making pseudo-labels to make the network detect the regions of object aggregation. The results are shown in Table V, although the way of QueryDet brings more runtime and complexity improvement, the improvement is limited, and the performance decreases too much. Our method balances performance and complexity, improving model efficiency without reducing performance. we also find that using a large distance would keep many invalid proposals, which has more complexity. We set the distance r_a to 2 after a trade-off between accuracy and complexity.

(a) P_2 (b) P_3 (c) $P_{4\text{-}6}$

Fig. 7. Visualization of proposals from different layers. Because there are fewer objects with large scales, we combine the detection results from P_{4-6}.

TABLE VI
COMPARISON OF PERFORMANCE AND COMPLEXITY USING PROPOSALS FROM DIFFERENT LAYERS

Layer	AP	AP_s	AP_m	AP_l
P_2	30.3	26.7	40.8	4.7
P_3	5.2	0.1	6.7	41.1
P_{4-6}	1.3	0	0.4	15.7

TABLE VII
COMPARISON OF PERFORMANCE AND COMPLEXITY USING DIFFERENT INITIAL PROPOSAL BOX NUMBERS

Num	AP	AP_s	AP_m	AP_l	s/img	FLOPs
1500	33.0	25.9	42.1	42.7	0.183	136.1G
2000	33.8	26.8	42.9	45.8	0.216	175.1G
2500	33.6	26.3	43.1	44.0	0.253	214.0G

TABLE VIII
COMPARISON OF THE DIFFERENT STAGE NUMBERS SDPDET USES

Num	AP	AP_s	AP_m	AP_l	s/img	FLOPs
2	14.5	10.6	19.3	19.9	0.145	71.2G
3	23.4	16.7	31.8	36.9	0.162	97.2G
4	30.9	23.6	40.1	43.1	0.179	123.2G
5	33.1	26.0	42.2	45.9	0.198	149.1G
6	33.8	26.8	42.9	45.8	0.216	175.1G
12	32.4	25.6	41.8	44.8	0.310	330.9G

D. Visualization of Proposals From Different Layers

For a further representation of the performance of our proposed SLP on scale separation, we visualize the detection results of proposals from different layers separately. The results are shown in Fig. 7, the proposals from P_2 only detect the small objects, while the proposals from P_3 and P_{4-6} detect only larger objects. This demonstrates that our proposed SLP makes learnable proposals scale-separated well, which allows specific proposals to focus only on specific scales.

E. Influence of Proposals From Different Layers

As described in Section IV-B, we explore the influence of proposals from each level on the detection ability. As shown in Table VI, when only using the proposals with the P_2, we get superior performance only on the AP_s and AP_m. With more layers involved (i.e. , from P_3 to P_{4-6}), the performance of AP_l is significantly improved (i.e. , from 5.3 to 41.1 and 15.7). It demonstrates that enabling the proposals from different layers to learn objects with different scales effectively facilitates the detection of multi-scale objects.

F. Influence of Proposals Numbers

The number of proposals pre-defined by the model limits the maximum number of objects predicted. To study the influence of proposal numbers on SDPDet, we progressively increase initial proposals from 1500 to 2500. As shown in Table VII, increasing the number of proposals to 2000 can effectively improve the detection performance. However, when the number of proposals is larger, it can significantly increase computational costs and even reduce performance. It is because a large number of proposals will cause duplication between the predictions of proposals. We finally set up 2000 initial proposals after a trade-off between accuracy and complexity.

G. Influence of the Number of Stages

Existing works have demonstrated that iterative structures [24], [46] can effectively improve object detection performance. As shown in Table VIII, with the number of stages increasing, the accuracy improves very significantly in the first few stages. However, further increasing the number of stages, the performance improvement is not obvious or even worse, which is because too many stages tend to result in accurate predictions being filtered out. Besides, the runtime and FLOPs are almost linearly related to the number of stages. After a trade-off between complexity and accuracy, we set six stages.

H. Influence of Different Label Assignments

To further validate the effectiveness of our method, we applied different label assignments on Sparse R-CNN and the results are

TABLE IX
COMPARISON OF THE DIFFERENT LABEL ASSIGNMENTS

Method	AP	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_l
IoU	7.8	13.3	7.9	7.0	10.3	8.0
GIoU	9.5	16.9	9.3	8.3	11.8	13.9
L2	-	-	-	-	-	-
IoU+cls	28.9	49.7	29.2	22.3	35.6	38.3
baseline	29.9	50.9	30.5	23.4	38.4	45.9
POTO [56]	32.3	55.3	32.7	25.5	40.9	40.9
Ours	33.7	56.6	34.3	26.7	42.9	45.7

IoU indicates selecting the proposal with the maximum IoU of the GT as the positive sample, GIoU indicates selecting the proposal with the maximum GIoU of the GT as the positive sample, L2 indicates selecting the proposal with the minimum L2 distance of the center of GT as the positive sample, and IoU+cls means combining the classification predictions of the proposals and the iou to select positive samples. '-' indicates that the model cannot converge.

shown in Table IX. IoU achieves a poor performance result (7.8 AP), GIoU also performs badly (9.5 AP) and L2 can't even converge. The static label assignment provides sub-optimal positive samples, and the few and inaccurate positive samples lead to inefficient feature learning in Sparse R-CNN due to its one-to-one matching strategy. With the supervision of classification added (iou+class), the performance improves significantly (+21.1 AP), which shows that joint supervision of classification and localization is vitally important for dynamic one-to-one label assignment in Sparse R-CNN. We also introduced another one-to-one label assignment from POTO [56]. Compared with the baseline, POTO has some improvement on small and medium objects (+2.1 AP_s, +2.5 AP_m), but has a larger performance gap on large objects (-5.0 AP_l). Since there are more small and medium objects in the dataset, the AP is also higher than the baseline. Our method improves the performance of small and medium objects while keeping the performance of baseline on large objects and achieves better performance than POTO, which shows the efficiency of our method.

VII. LIMITATIONS

Although SDPDet can effectively localize objects with discrete distributions and various scales, it still suffers from some limitations. 1) Compared with the two-step detection methods of fine detection by zooming in sub-regions, the appearance features of objects used in SDPDet are very limited, which reduces the classification performance. 2) When training the detector in the two-step detection methods using the zoomed-in sub-regions, it implicitly expands the data magnitude and optimizes the class distribution. In contrast, SDPDet also suffers from extreme class distributions (e.g., long-tail).

VIII. CONCLUSION

In this paper, we pay more attention to the uneven distribution and scale variation in drone-view images and present a new end-to-end drone-view image detection model, called SDPDet. In particular, SDPDet includes 1) a scale-separated activation pyramid (SAP) to focus on the regions that have object aggregation at each scale, and 2) a scale-separated learnable proposals (SLP) mechanism to learn proposals in these regions at different scales. To our best knowledge, SDPDet is the first work to enable the learnable proposals mechanism to handle object uneven distribution and scale variation in drone-view object detection. The experimental results show a significant superiority between the proposed SDPDet and the existing models. In the future, we will enable the scale-separated learnable proposals mechanism to solve the problem of unbalanced category distribution through more explorations.

REFERENCES

[1] X.-W. Tang, X.-L. Huang, and F. Hu, "QoE-driven UAV-enabled pseudo-analog wireless video broadcast: A joint optimization of power and trajectory," *IEEE Trans. Multimedia*, vol. 23, pp. 2398–2412, 2020.

[2] S. Zhang et al., "Person re-identification in aerial imagery," *IEEE Trans. Multimedia*, vol. 23, pp. 281–291, 2021.

[3] F. Zhang et al., "Image-only real-time incremental UAV image mosaic for multi-strip flight," *IEEE Trans. Multimedia*, vol. 23, pp. 1410–1425, 2021.

[4] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[5] M. Everingham, L. V. 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.

[6] S. Shao et al., "Objects365: A large-scale, high-quality dataset for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8429–8438.

[7] X. Li, S. Lai, and X. Qian, "DBCFace: Towards pure convolutional neural network face detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1792–1804, Apr. 2022.

[8] C. Liu et al., "Product recognition for unmanned vending machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1584–1597, Feb. 2024.

[9] H. Jin, S. Lai, Q. Tang, T. Zhu, and X. Qian, "MPPM: A mobile-efficient part model for object re-ID," *IEEE Trans. Multimedia*, vol. 25, pp. 6356–6370, 2023.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[12] P. Zhu et al., "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2022.

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

[14] 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. 8310–8319.

[15] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 737–746.

[16] S. Deng et al., "A global-local self-adaptive network for drone-view object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1556–1569, 2020.

[17] C. Duan, Z. Wei, C. Zhang, S. Qu, and H. Wang, "Coarse-grained density map guided object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 2789–2798.

[18] J. Xu, Y. Li, and S. Wang, "AdaZoom: Towards scale-aware large scene object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 4598–4609, 2023.

[19] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[20] 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.

[21] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13658–13667.

[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[23] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[24] P. Sun et al., "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14454–14463.
[25] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke
[26] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9756–9765.
[27] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 303–312.
[28] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor DETR: Query design for transformer-based detector," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2567–2575.
[29] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[30] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[31] 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. 2999–3007.
[32] K. Duan et al., "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6568–6577.
[33] 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. 9626–9635.
[34] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, *arXiv:1902.07296*.
[35] C. Chen et al., "RRNet: A hybrid detector for object detection in drone-captured images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 100–108.
[36] J. Li et al., "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1951–1959.
[37] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 206–221.
[38] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9724–9733.
[39] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 845–853.
[40] J.-S. Lim, M. Astrid, H.-J. Yoon, and S.-I. Lee, "Small object detection using context and attention," in *Proc. Int. Conf. Artif. Intell. Inf. Commun.*, 2021, pp. 181–186.
[41] K. Fu, J. Li, L. Ma, K. Mu, and Y. Tian, "Intrinsic relationship reasoning for small object detection," 2020, *arXiv:2009.00833*.
[42] W. Shen, P. Qin, and J. Zeng, "An indoor crowd detection network framework based on feature aggregation module and hybrid attention selection module," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 82–90.
[43] T.-Y. Lin et al., "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
[44] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.
[45] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
[46] Q. Hong et al., "Dynamic sparse R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4713–4722.
[47] S. Contributors, "SpConv: Spatially sparse convolution library," 2022. [Online]. Available: https://github.com/traveller59/spconv
[48] H. Rezatofighi et al., "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 658–666.
[49] Y. Wang, Y. Yang, and X. Zhao, "Object detection using clustering algorithm adaptive searching regions in aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 651–664.
[50] J. Liao et al., "Unsupervised cluster guided object detection in aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11204–11216, 2021.
[51] Y. Huang, J. Chen, and D. Huang, "UFPMP-Det: Toward accurate and efficient object detection on drone imagery," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 1026–1033.
[52] J. Leng et al., "Pareto refocusing for drone-view object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1320–1334, Mar. 2022.
[53] A. Meethal, E. Granger, and M. Pedersoli, "Cascaded zoom-in detector for high resolution aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2023, pp. 2046–2055.
[54] O. C. Koyun, R. K. Keser, İ. 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.
[55] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
[56] J. Wang et al., "End-to-end object detection with fully convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15844–15853.

Nengzhong Yin received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 2022, and the M.E. degree from SMILES Lab, Xi'an Jiaotong University, Xi'an, China, in 2025. His research interests include object detection and object tracking.

Chengxu Liu (Graduate Student Member, IEEE) received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 2019. He is currently working toward the Ph.D. degree with SMILES Lab, Xi'an Jiaotong University, Xi'an, China. From 2021 to 2022, he was an Intern with Multimedia Search and Mining Group, Microsoft Research Asia, and with Invictus Sense Group, MEGVII Research from 2022 to 2024. He is also a Visiting Research Student with Vision and Learning Lab, University of California, Merced, CA, USA. His research interests include low-level vision, object detection, and recognition.

Ruhao Tian is currently working toward the B.E. degree from Xi'an Jiaotong University, Xi'an, China. His research interests include brain-inspired computing and brain-inspired spiking neural networks.

Xueming Qian received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in electronics and Information engineering from Xi'an Jiaotong University, Xi'an, China, in 2008. From 2011 to 2014, he was an Associate Professor with Xi'an Jiaotong University, where he is currently a Full Professor and the Director of SMILES Lab. From 2010 to 2011, he was a Visiting Scholar with Microsoft Research Asia, Beijing, China. His research interests include social media Big Data mining and search. Prof. Qian was the recipient of the Microsoft Fellowship in 2006 and the Outstanding Doctoral Dissertations of Xi'an Jiaotong University and Shaanxi Province in 2010 and 2011, respectively.