# AD-Det: Boosting Object Detection in UAV Images with Focused Small Objects and Balanced Tail Classes

**Authors:** Zhenteng Li

**Date:** 2025-04-06T00:00:00+00:00

## Abstract

Object detection in Unmanned Aerial Vehicle (UAV) images poses significant challenges due to complex scale variations and class imbalance among objects. Existing methods often address these challenges separately, overlooking the intricate nature of UAV images and the potential synergy between them. In response, this paper proposes AD-Det, a novel framework employing a coherent coarse-to-fine strategy that seamlessly integrates two pivotal components: Adaptive Small Object Enhancement (ASOE) and Dynamic Class-balanced Copy-paste (DCC). ASOE utilizes a high-resolution feature map to identify and cluster regions containing small objects. These regions are subsequently enlarged and processed by a fine-grained detector. On the other hand, DCC conducts object-level resampling by dynamically pasting tail classes around the cluster centers obtained by ASOE, maintaining a dynamic memory bank for each tail class. This approach enables AD-Det to not only extract regions with small objects for precise detection but also dynamically perform reasonable resampling for tail-class objects. Consequently, AD-Det enhances the overall detection performance by addressing the challenges of scale variations and class imbalance in UAV images through a synergistic and adaptive framework. We extensively evaluate our approach on two public datasets, i.e., VisDrone and UAVDT, and demonstrate that AD-Det significantly outperforms existing competitive alternatives. Notably, AD-Det achieves a 37.5% Average Precision (AP) on the VisDrone dataset, surpassing its counterparts by at least 3.1%.

## Full Text

## Preamble

**AD-Det: Boosting Object Detection in UAV Images with Focused Small Objects and Balanced Tail Classes**

Zhenteng Li[1], Sheng Lian[1,2,*], Dengfeng Pan[1], Youlin Wang[1], and Wei Liu[3]

[1]College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China

Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350002, China

School of Software, East China Jiaotong University, Nanchang 330013, China

*Correspondence: S. Lian: shenglian@fzu.edu.cn

**Abstract:** Object detection in Unmanned Aerial Vehicle (UAV) images poses significant challenges due to complex scale variations and class imbalance among objects. Existing methods often address these challenges separately, overlooking the intricate nature of UAV images and the potential synergy between them. In response, this paper proposes AD-Det, a novel framework employing a coherent coarse-to-fine strategy that seamlessly integrates two pivotal components: Adaptive Small Object Enhancement (ASOE) and Dynamic Class-balanced Copy-paste (DCC). ASOE utilizes a high-resolution feature map to identify and cluster regions containing small objects. These regions are subsequently enlarged and processed by a fine-grained detector. On the other hand, DCC conducts object-level resampling by dynamically pasting tail classes around the cluster centers obtained by ASOE, maintaining a dynamic memory bank for each tail class. This approach enables AD-Det to not only extract regions with small objects for precise detection but also dynamically perform reasonable resampling for tail-class objects. Consequently, AD-Det enhances the overall detection performance by addressing the challenges of scale variations and class imbalance in UAV images through a synergistic and adaptive framework. We extensively evaluate our approach on two public datasets, i.e., VisDrone and UAVDT, and demonstrate that AD-Det significantly outperforms existing competitive alternatives. Notably, AD-Det achieves a 37.5% Average Precision (AP) on the VisDrone dataset, surpassing its counterparts by at least 3.1%. Code will be available at https://github.com/gentlezachary/AD-Det.

**Keywords:** object detection; UAV images; scale variations; class imbalance

## 1. Introduction

Unmanned Aerial Vehicles (UAVs) have experienced widespread adoption in recent years, driven by their cost-effectiveness and adaptability. Outfitted with cameras and various sensors, UAVs find application in diverse scenarios such as surveillance, rescue operations, agriculture, express delivery, and more. The effectiveness of these applications relies heavily on the seamless and robust recognition of objects within the UAV's field of view, as noted in recent comprehensive reviews [?, ?, ?].

Solutions including R-CNN series [?, ?], YOLO series [?, ?] and most recent Query-based series [?, ?, ?] have shown leading performance on public datasets such as MS COCO [?]. Despite these accomplishments, when applied to UAV images, particularly datasets like VisDrone [?], generic detectors' performance falls short of satisfaction in terms of accuracy and efficiency. Challenges arise due to hardware limitations, complexities in the imaging environment, and flight trajectory intricacies, leading to the following issues:

**Scale variations:** Owing to variations in flight altitudes and complex object distribution in UAV images, object sizes differ by distance from camera exposure, exhibiting a broad range of scales, with a predominance of small-scale objects. As illustrated in Figure 1 Figure 1: see original paper, within the VisDrone dataset, the proportion of small objects (object size $< 32^2$ pixels) reaches 60.5%, which is 19.1% higher than that in the COCO dataset. On the other hand, the percentage of large objects (object size $> 96^2$ pixels) accounts for only 5.5%. The limited feature representation of small objects poses a considerable challenge, which leads to decreased accuracy and reliability in detecting smaller objects, thus significantly affecting the overall detection performance.

**Class imbalance:** UAV images collected in urban environments exhibit a long-tailed class imbalance characteristic: a small subset of head classes (e.g., car and people) dominates the dataset, collectively representing over 70% of object instances, while numerous tail classes (e.g., bicycle and van) are severely under-represented, constituting only 1% to 7% of instances (Figure 1(b)). This skewed distribution creates a model bias towards head classes during training—detectors prioritize optimizing features for high-frequency categories at the expense of tail classes. Consequently, tail classes suffer from poor localization accuracy and high false-negative rates due to insufficient discriminative feature learning.

To address these challenges, researchers have made many attempts. For scale variations, [?, ?] focus on multi-scale feature fusion, [?, ?] emphasize hard sample mining. Nevertheless, due to the input scale limitation, there is still room for improvement in accuracy. Another direct way is to crop the image into subregions before applying detection, such as uniform cropping. However, these cropping strategies cannot adaptively adjust based on the semantic information in UAV images, potentially including extensive background areas. Li et al. [?] utilize the density map to guide the cropping of images, enhancing the semantics of the resulting subregions. However, this approach contains a density map generation network, which increases model complexity and requires ground truth density map generation. Compared to the scale variations, the challenge of class imbalance is often neglected in UAV images. In recent years, research in the field of long-tailed detection has primarily been categorized into three types: (1) resampling and data augmentation techniques [?]; (2) loss function reweighting methods [?, ?]; and (3) decoupled optimization strategies [?]. However, existing approaches still face challenges in dynamically adapting to changes in the distribution of tail classes. Zhang et al. [?] employ a Multi-Model Fusion (MMF) strategy to tackle the head and tail classes distinctly, effectively enhancing the detection performance of tail classes. However, this approach results in the discarding of a substantial amount of valuable data during the training of each model, which could potentially compromise the model' s representational capacity.

In this paper, for object detection in UAV images, we propose a novel framework called AD-Det, which adopts a coarse-to-fine strategy and mainly consists of two key components: Adaptive Small Object Enhancement (ASOE) and Dynamic Class-balanced Copy-paste (DCC). Specifically, ASOE employs a high-resolution feature map from the classification head to pinpoint small objects. As shown in Figure 1(c), most small objects in the image show significant activation in the high-resolution feature map. These positions of interest are clustered into N subregions, which are then enlarged and processed by a fine-grained detector, thereby enhancing the detection performance of small objects.

DCC, on the other hand, performs object-level resampling through a dynamic copy-and-paste strategy specifically tailored for tail-class objects. It leverages the clustering cues from ASOE to dynamically search for suitable pasting positions around the cluster center, and maintain a dynamic memory bank for each tail class. Additionally, data augmentation is conducted to avoid overfitting. In this way, the proposed AD-Det can extract regions with small objects for fine-grained detection and dynamically perform reasonable resampling for tail-class objects, thereby improving overall detection performance.

Compared to existing coarse-to-fine solutions, our approach is a plug-and-play, unsupervised scheme. The advantages lie in leveraging the inherent properties of the detector to focus on small object regions that truly contribute to accuracy gains, thereby enhancing the precision of small object detection without extra learnable parameters. Additionally, we propose a training-only image augmentation technique for tail classes, which helps alleviate class imbalance issues.

We summarize four main contributions of this work as follows:

- We propose AD-Det, a novel object detection framework that addresses two key challenges in UAV image object detection: complex scale variations and class imbalance, ultimately enhancing the detection performance.
- To enhance small object detection, we propose ASOE, which uses a high-resolution feature map to cluster regions containing small objects, followed by processing them for fine-grained detection.
- To enhance tail-class object detection, we propose DCC, which performs reasonable object-level resampling by dynamically pasting tail classes around clustering centers obtained by ASOE.
- We conduct extensive experiments on VisDrone and UAVDT, which demonstrate that our approach significantly outperforms existing competitive alternatives.

## 2. Related Works

In object detection tasks, unlike traditional solutions that rely on handcrafted features and classifiers, DL-based solutions can automatically learn discriminative features and achieve promising results. Here, we briefly review object detection solutions in the deep learning era, including (1) Generic object detection (in Section 2.1) that focus on natural images such as MS COCO [?], (2) Object detection in aerial images (in Section 2.2) that focus on aerial images such as VisDrone [?], DOTA [?] and (3) Long-tail object detection (in Section 2.3) that focus on long-tail scenarios such as LVIS [?].

### 2.1. Generic object detection

Sparked by the impressive success in image classification tasks [?], DL-based methods have dominated in generic object detection. The existing DL-based detection solutions, when analyzed according to their pipeline, can be roughly grouped into two-stage and one-stage methods. The two-stage detectors are represented by R-CNN [?], Fast R-CNN [?], Faster R-CNN [?] and Mask R-CNN [?]. These detectors first reduce the search space significantly by extracting candidate region proposals. The extracted proposals are then classified into specific categories and refined to proper locations. Such a pipeline leads to relatively higher accuracy but lower efficiency. Different from the above methods, one-stage methods, on the other hand, directly predict objects' categories and locations without region proposals. The representative approaches include YOLO series [?, ?, ?], SSD [?], and RetinaNet [?]. Such design leads to relatively higher efficiency at the sacrifice of accuracy. Recently, researches such as CornerNet [?] and FCOS [?] bypassed the anchor boxes mechanism and corresponding hyper-parameter setting, and gave promising alternatives for one-stage methods. DETR [?] pioneered the fully end-to-end object detector that employed a transformer-based architecture, eliminating the reliance on anchor generation and non-maximum suppression (NMS).

Although these detectors achieved impressive progress in generic object detection, the performance in UAV images is far from satisfying due to problems of scale variations and class imbalance.

### 2.2. Object detection in aerial images

Compared with generic objects that are mostly captured in ground view, object detection in UAV images presents heightened challenges due to object scale and object/category distribution problems.

Many studies focus on multi-scale/tiny-scale object problems in aerial images. Most of the early research is the migration of classical generic object detection algorithms. Yang et al. [?] proposed a novel query mechanism, which effectively and efficiently incorporates an additional high-resolution layer to promote small object detection. In the realm of knowledge distillation, Zhu et al. [?] costlessly enhanced the performance of lightweight models by incorporating scale-aware knowledge from more complex ones. To detect small weak objects in UAV images, Han et al. [?] presented a context-scale-aware detector, combining the strengths of context-aware learning and multi-scale feature extraction. Cao et al. [?] introduced a self-reconstructed difference map approach to enhance feature visibility for challenging tiny object detection tasks. DQ-DETR [?] specialized in tiny object detection by employing dynamic query selection and counting-guided feature enhancement, boosting detection performance. SDPDet [?] employed scale-separated dynamic proposals and activation pyramids to enhance the efficiency and accuracy of object detection in UAV views. Zhang et al. [?] enabled detectors to focus on discriminative features while reducing false positives in cluttered backgrounds. Zhang et al. [?] proposes IMPR-Det, which can integrally mix multiscale pyramid representations for different components of an instance.

Besides object scale problems, many studies focus on object/category distribution problems. Inspired by the research of crowd counting, Li et al. [?] injected density maps estimation into the conventional object detection framework, which was utilized to predict object distribution, reduce the influence of background, and generate more balanced candidate proposals. Duan et al. [?] further adopted a coarse-grained density map to identify subregions more accurately. Similar observations regarding the object distribution issues led [?] to improve RPN to cluster region proposals. GLSAN [?] utilized a scale-aware algorithm to fuse global and local detection results. CZDet [?] enhanced cascade detection by innovatively incorporating high-density subregion labels into the repurposed detector. YOLC [?] employed a Local Scale Module and deformable convolutions to enhance accuracy in aerial small object detection. Zhang [?] proposed a structured adversarial self-supervised pretraining to strengthen both clean accuracy and adversarial robustness. On the other hand, Yu et al. [?] investigated the long-tail category distribution issues, where they designed dedicated samplers and detection heads to address the distinct characteristics of tail and head classes.

Such excellent works indeed facilitate aerial image object detection. Nevertheless, when addressing scale variations and class imbalance challenges, the majority of preceding studies treat these essential issues rigidly and separately, not only ignoring the complexity of UAV images but also neglecting their potential synergy.

### 2.3. Long-tail object detection

Much like the paradigm of long-tail classification, research in the domain of long-tail object detection predominantly adheres to two distinctive methodologies: resampling and reweighting. Commonly employed techniques in resampling involve either oversampling the minority classes or undersampling the majority classes during the training process, aimed at mitigating the issue of imbalanced class distribution. Repeat factor sampling (RFS) [?] presented an image-level sampling approach following the resampling paradigm. SimCal [?] proposed a bi-level class balanced sampling approach to alleviate classification head bias. In the context of reweighting, the fundamental concept lies in the assignment of diverse weights to training samples, with a focus on enhancing the training of tail samples. Cui et al. [?] introduced a novel loss function that significantly improves class imbalance handling by dynamically adjusting the weights of training samples based on their effective numbers.

Besides the aforementioned strategies, many researchers have endeavored to handle the long-tail problem from other perspectives. Kang et al. [?] advocated a decoupled training paradigm that disentangles the learning process into distinct phases: representation learning and classifier training. BAGS [?] grouped classes based on their instance frequency and subsequently applied softmax loss among each group. ROG [?] proposed a multi-task learning approach that concurrently optimizes both object-level classification and global-level score ranking. Hyun et al. [?] introduced an Effective Class-Margin Loss (ECM) as a novel surrogate objective to optimize margin-based binary classification error on the imbalanced training set.

Many earlier coarse-to-fine counterparts pinpoint regions of interest by relying on extra learnable modules or by post-processing sparse boxes. Our method, however, differs by incorporating dense indicators from a high-resolution feature map. This map is adept at preserving extensive information about small objects, facilitating adaptive region generation. Additionally, we enhance this framework by addressing class imbalance issues through the integration of object-level copy-paste techniques. This novel combination effectively tackles two distinct challenges, which are overlooked in existing solutions.

## 3. Methodology

In this section, we elaborate on the design of AD-Det, including the overall framework of AD-Det in Section 3.1, the ASOE module in Section 3.2, the DCC module in Section 3.3, along with the training and inference details of AD-Det in Section 3.4.

Figure 2 [Figure 2: see original paper] illustrates the framework overview of the proposed AD-Det. The network adopts a coarse-to-fine strategy and integrates two key modules: an Adaptive Small Object Enhancement (ASOE) module for excavating regions containing small objects and a dynamic Class-balanced Copy-paste (DCC) module for balancing class distribution. The final results are obtained by fusing global results and local results using non-maximum suppression (NMS). The dashed line between ASOE and DCC indicates training-only operation.

### 3.1. Framework overview

To tackle the key challenges of object detection in UAV images—scale variations and class imbalance—we propose a novel approach called AD-Det. Designed in a coarse-to-fine manner, AD-Det mainly consists of two key components. Adaptive Small Object Enhancement (ASOE) utilizes coarse detector cues to roughly locate small objects and cluster regions of interest for later fine-grained detection.

Dynamic Class-balanced Copy-paste (DCC) performs reasonable object-level resampling by dynamically pasting tail classes around the cluster centers obtained by ASOE. The predictions given by the coarse-to-fine strategy are fused through non-maximum suppression (NMS).

### 3.2. Adaptive Small Object Enhancement (ASOE)

In UAV images, small objects constitute a large proportion and are sparsely distributed (Figure 1(a,c)), resulting in unsatisfactory detection performance. Such small objects are typically detected from high-resolution low-level feature maps, for example, P3 in FPN [**?**]. As shown in Figure 3 [Figure 3: see original paper], the image is passed through the backbone to extract multi-scale features, which are subsequently fused in the FPN. These fused features are then passed to the detection head. Due to varying convolutional strides in different layers, the feature map sizes decrease from P3 to P5. The lower layers, having undergone fewer downsampling operations, not only retain more local information but also have a receptive field better aligned with small objects. As observed from the heatmap comparison, small objects are primarily detected at the P3 layer. Specifically, the human visual system excels at object detection by quickly scanning the entire image to acquire information on large objects and gain valuable clues in challenging areas with small objects.

Inspired by these factors, we propose Adaptive Small Object Enhancement (ASOE), which utilizes a high-resolution feature map to pinpoint regions with small, difficult-to-detect objects for fine-grained detection. Unlike [**?**], which relies on an additional density network to obtain the position of foreground objects, the proposed ASOE module employs a hierarchical approach, seamlessly integrating coarse-to-fine strategies to overcome the inherent challenges of identifying small objects in UAV images.

Figure 4 [Figure 4: see original paper] illustrates the process of subregion generation in ASOE. In the first stage, a detector Coarse( $\cdot$ ) captures the global context and produces preliminary detection results. Meanwhile, the feature map $F_l \in \mathbb{R}^{C\times H\times W}$ (Figure 2 coarse stage red layer), targeting semantic details pertinent to small objects, is extracted from the classification head enriched with high-resolution and low-level features (Figure 3), akin to the methodology employed by RetinaNet [**?**]. Here, $l$ represents the layer index of $P_l$, while $C$, $H$, and $W$ represent the channel count, height, and width of the feature map. As shown in Figure 4, we compute a position-wise activation map $V_l \in \mathbb{R}^{1\times H\times W}$ by applying the Sigmoid function independently to each channel of $F_l$, followed by averaging across channels: $\sigma(F_l)$, where $V_l(i,j)$ indicates the probability that the grid $(i,j)$ contains a small object and $\sigma$ is the Sigmoid function. To suppress background noise, we retain only positions where $V_l(i,j)$ exceeds a threshold $\gamma = 0.5$, and by multiplying with the downsampling factor $2^l$, these positions are mapped to the corresponding locations in the original image:

$$T_l = \{(i \times 2^l, j \times 2^l) \mid V_l(i,j) > \gamma\}$$

These identified positions approximate the small objects' locations, which are prone to being overlooked or inaccurately estimated. To address the non-uniform distribution of small objects in UAV images, the ASOE module utilizes the K-means algorithm to cluster $T_l$ into $N$ distinct clusters. The K-means algorithm achieves effective performance gains through K-value optimization, requiring only minimal parameter adaptation for cross-dataset deployment. This balances deterministic region control with adaptability to diverse data distributions, avoiding the parameter sensitivity of DBSCAN and computational overhead of mean-shift for real-time UAV applications. The cluster process can be represented as $\{C_1, \ldots, C_N\} = \text{K-means}(T_l, N)$, where $C_k = \{(i,j) \in T_l \mid k = \arg\min_m \|x - \mu_m\|^2\}$. By determining the top-left and bottom-right coordinates of each cluster, we crop $N$ subregions from the original image, defined as $\{S_i, \ldots, S_N\} = \text{CropByTLBR}(C)$. Subsequently, we upscale the extracted $N$ subregions and input them into a fine-grained detector, thereby enhancing the detection performance of small objects.

### 3.3. Dynamic Class-balanced Copy-paste (DCC)

To address the challenge of class imbalance in UAV images, resampling techniques like class-balanced sampling and repeat factor sampling are commonly employed. However, these image-level resampling solutions elevate training costs. In contrast, the copy-paste method offers object-level resampling by copying and pasting objects, which has significantly advanced object detection in ordinary images. However, its application to UAV images has been less effective, attributed to their complexity. Therefore, we propose Dynamic Class-balanced Copy-paste (DCC), which differs from traditional copy-paste solutions by considering both the position rationality and diversity of objects.

Firstly, to make the pasting position more reasonable, DCC utilizes a dynamic search (DS) strategy for suitable positions to paste by combining ASOE clustering cues. Specifically, in ASOE, besides obtaining subregions for fine-grained detection, we can also get their clustering centers, which represent the distribution centers of objects in each subregion. DCC starts from the clustering centers and uses the BFS algorithm [?] to dynamically search for suitable positions around them. Here, "Suitable" denotes positions that are proximal and maintain no overlap with adjacent objects. To mitigate overfitting caused by repetitive resampling of specific instances, we introduce a Diversity Augmentation (DA) strategy, leveraging dynamic memory banks with data augmentation techniques to ensure diverse and balanced training batches. Each tail category is assigned a memory bank with a capacity of 10, dynamically updated using a first-in-first-out (FIFO) replacement policy during subregion traversal. Before pasting, we introduce data augmentation techniques such as Shift Scale Rotate and Random Brightness Contrast.

Algorithm 1 details the generation of fine-grained subregions. The algorithm takes as input a normalized high-resolution heatmap $V$, threshold $\gamma$, the number of subregions $N$, and the training state. It outputs subregions $S$ by first filtering positions based on the threshold, then applying K-means clustering, and finally cropping subregions. During training, the DCC module is applied to augment tail-class instances.

### 3.4. Training and inference details

In the training process, AD-Det integrates ASOE and DCC to independently train the coarse and fine-grained detectors. To achieve coarse detection and extract the small object feature map, the coarse detector is first trained using all original images, with its training loss function defined as:

$$L_{\text{coarse}} = L_{\text{cls}}(x_g, y_g^{\text{cls}}) + L_{\text{reg}}(x_g, y_g^{\text{reg}})$$

where $L_{\text{cls}}$ and $L_{\text{reg}}$ denote classification loss and regression loss, and $\{x_g, y_g\}$ denote the original training set.

Then, we conduct the coarse detector to traverse the original training set. For each sample, ASOE extracts small object subregions and combines with DCC to augment tail classes. This process yields the fine-grained training set $\{x_l, y_l\}$, the detailed algorithm is illustrated in Algorithm 1, and the training loss function for the fine-grained detector can be represented as:

$$L_{\text{fine}} = L_{\text{cls}}(x_l, y_l^{\text{cls}}) + L_{\text{reg}}(x_l, y_l^{\text{reg}})$$

In the inference process, due to the fine-grained detector's capability of detecting subregions with small objects and long-tail distributions, AD-Det exclusively relies on ASOE to acquire related subregions. Let $G$ denote the input UAV image, and the final detection result $D_{\text{final}}$ can be formulated as:

$$D_{\text{final}} = \text{NMS}\left(\text{Coarse}(G), \bigcup_i \text{Fine}(S_i)\right)$$

where Coarse( $\cdot$ ) and Fine( $\cdot$ ) denote coarse and fine-grained detectors, respectively. $S_i$ is the subregions extracted by ASOE, and NMS stands for non-maximum suppression operation.

## 4. Experiments

### 4.1. Experimental setup

**Datasets.** We evaluate our AD-Det on two publicly available datasets for UAV image object detection: VisDrone [?] and UAVDT [?]. We summarize the details on these two datasets in Table 1 .

VisDrone is a challenging large-scale dataset captured by various camera devices using multiple UAVs. This dataset contains data collected from several Chinese cities, covering different weather conditions and scenarios. Ten predefined categories, including pedestrian, car, van, etc., are manually annotated with bounding boxes. There are 6,471 images in the VisDrone for training, complemented by 548 images in the validation set, and a total of 3,190 images for testing. The maximal resolution for each image is 2000 × 1500 pixels. Given that the testing set is not publicly available, we follow ClusDet [?] and DMNet [?] to train the model on the training set while evaluating it on the validation set.

UAVDT is a large-scale vehicle detection and tracking dataset for UAV scenarios, which is released by UCAS. This dataset contains data collected from complex environments with different weather conditions, occlusion, and flying heights. Bounding boxes have been manually annotated for three predefined categories, including car, truck, and bus. It contains 23,258 and 15,069 images for training and testing, respectively, and the resolution is 1080 × 540 pixels for each image.

**Evaluation metrics.** Following the evaluation protocols outlined in MS COCO [?], we employ Average Precision (AP) as our primary metric, spanning various categories and IoU thresholds. The metrics are succinctly outlined as follows:

- **AP:** Average Precision calculated across all categories, considering IoU values within the range of [0.5, 0.95] at intervals of 0.05.
- **AP_{50}, AP_{75}:** Average Precision computed across all categories using individual IoU thresholds of 0.5 and 0.75, respectively.
- **AP_S, AP_M, AP_L:** Average Precision calculated on object size of small (less than $32^2$), medium (from $32^2$ to $96^2$), and large (greater than $96^2$).

**Implementation details.** We implement our AD-Det using the MMDetection toolbox [?] with PyTorch [?] as the basic framework. The model is trained on a system equipped with 2×Intel Silver 4210R CPU and 2×NVIDIA A4000 GPU. The baseline detection network employed is GFL [?] with 4 parts uniform cropping.

**Training Phase:** For the training procedure on VisDrone and UAVDT, as [?] does, the input image size is configured as 1333 × 800 pixels for VisDrone and 1024 × 540 pixels for UAVDT, with both coarse and fine-grained detectors. Concerning hyperparameters in the ASOE, the maximum subregion number ($N$) is set to 4 and 3 on VisDrone and UAVDT, respectively (discussed in Section 4.4). In the DCC module, all categories except pedestrian, people, and car are considered as tail classes on VisDrone, while truck and bus are considered as tail classes on UAVDT. All models undergo 12 epochs of training using the SGD optimizer, with momentum = 0.9, weight decay = 0.0001, and initial learning rate = 0.01. AD-Det is trained with a linear warm-up strategy and undergoes decay by a factor of 10 at epochs 8 and 11.

**Testing Phase:** The input image size and hyperparameters in the ASOE remain consistent with the training phase unless otherwise specified. In the process of fusing detection, the non-max suppression (NMS) threshold and max detection number are respectively set to 0.5 and 500 across all datasets.

### 4.2. Comparison with representative solutions

**Results on VisDrone.** We compare AD-Det with other state-of-the-art solutions on VisDrone, and the results on AP, AP_{50}, and AP_{75} are listed in Table 2 . Our baseline method employs GFL with ResNet-50/101 and ResNeXt-101 as backbones. From the table, we have the following observations. (1) Our proposed method achieves consistent improvement across all compared solutions. Our model promotes detection performance to 37.5%, 60.9%, and 39.2% in terms of AP, AP_{50}, and AP_{75}, surpassing all compared methods. Among the compared solutions, ClusDet, DMNet, CDMNet, GLSAN, CZDet, AMRNet, and YOLC employ a coarse-to-fine strategy, which is similar to our approach. Even compared to CZDet, the best solution within them, our model also increases performance by 3.1%, 1.2%, and 4.6% in AP, AP_{50}, and AP_{75}, respectively. (2) Despite using weaker backbones, our method surpasses or rivals the performance of all compared methods. As illustrated at the bottom of Table 2, although we employed a relatively weak, lightweight backbone ResNet-50, our model continues to be highly competitive. Our method exhibits substantial improvements in AP metric compared to other methods, even with stronger backbones. For example, compared to YOLC which used ResNeXt-101 as the backbone, our method, with ResNet-50 as the backbone, nonetheless managed to increase performance by 2.2%, 1.4%, and 3.4% in AP, AP_{50}, and AP_{75}, respectively.

To further evaluate the accuracy and computational complexity of AD-Det, we present a comparison with existing solutions based on AP, parameter counts (Params), floating-point operations (FLOPs), and inference speed (s/img). As shown in Table 3 , under the same backbone architecture of ResNeXt-101, AD-Det demonstrates superior improvements in AP, Params, FLOPs, and inference speed compared to ClusDet, DMNet, and GLSAN. Although our method exhibits marginally higher FLOPs and slightly slower inference speed than CZDet and YOLC, it achieves higher detection accuracy and a significant reduction in Params. These results highlight that our approach delivers competitive performance with a better accuracy-speed trade-off, fulfilling practical engineering requirements for real-world deployment.

**Results on UAVDT.** Similar situations occur when our method is evaluated on UAVDT, and the detection accuracy compared with state-of-the-art methods on the testing set of UAVDT is shown in Table 4 . To facilitate comparison, we utilize GFL with ResNet-50 backbone as our baseline. Compared with other state-of-the-art methods, our method attains a new state-of-the-art performance even with relatively weaker ResNet-50 as the backbone, achieving 20.1% in AP, 34.2% in AP_{50}, and 21.9% in AP_{75}. In comparison with GLSAN, our model demonstrates increases of 3.1%, 6.1%, and 3.1% in AP, AP_{50}, and

AP_{75}, respectively.

### 4.3. Qualitative analysis

To qualitatively contrast the performance, Figure 5 [Figure 5: see original paper] depicts the detection results for both the baseline detector and our AD-Det. It is evident that our method surpasses the baseline detector and reaches satisfactory detection accuracy. Specifically, as depicted in the dashed area of Figure 5 and the corresponding zoom-in views, our approach excels in detecting small objects and tail categories. This superiority can be attributed to the contributions of ASOE and DCC, which collectively enhance UAV image object detection from two distinct perspectives.

### 4.4. Ablation study

We conduct extensive studies on VisDrone to validate the effectiveness of our key design, including (1) the effectiveness of ASOE and DCC modules, (2) the choice of key hyperparameters, (3) the selection of base detector, (4) the design of ASOE module, and (5) the design of DCC module. For all ablation studies, we employ ResNet-50 as our backbone.

**The effectiveness of ASOE and DCC modules.** As the key components of our method, ASOE and DCC are systematically incorporated into the model to evaluate their respective efficacy. Simultaneously, we demonstrate the improvements resulting from the integration of ASOE and DCC, emphasizing their mutual complementarity. A comprehensive presentation of the ablation study is provided in Table 5 and Figure 6 [Figure 6: see original paper], which illustrates the accuracy improvements across different metrics after integrating our key components.

As illustrated in Table 5, the incorporation of ASOE results in a notable improvement of 2.2% in AP and 3.0% in AP_S over the baseline method. Similarly, DCC contributes to an enhancement in detection performance, elevating it from 35.3% to 35.9% in AP. Compared to baseline with identical parameter settings, our method achieves a notable gain in accuracy, but only increases a minor time cost.

As illustrated in Figure 6, a noteworthy observation is the consistent enhancement in results for objects in the tail categories, especially for bicycles, tricycles, awn., and bus, improved by 0.9%   1.9%, underscoring the ability of DCC to augment detection performance reliably. Significantly, the addition of complementary samples by DCC does not adversely affect the performance of other categories.

**The choice of key hyperparameters.** We analyze the impact of two key hyperparameters: the number of subregions $N$ in Figure 7(a) [Figure 7: see original paper] and the confidence threshold $\gamma$ in Figure 7(b).

The number of subregions $N$ is a pivotal parameter that significantly influences detection speed and accuracy. The effects of different values of $N$ on detection time and accuracy are summarized in Figure 7(a). This subfigure indicates that as the number of subregions increases, the network' s detection accuracy improves, but with an associated increase in detection time. Considering the trade-off between accuracy and efficiency, this study selects $N = 4$ as the optimal number of subregions. We conduct a similar experiment on UAVDT and find that $N = 3$ is the optimal choice for UAVDT.

The confidence threshold $\gamma$ determines the quality of the subregions (Eq. 2). We experiment with different $\gamma$ values in $\{0.1, 0.3, 0.5, 0.7\}$. In particular, we quantify sample points and their corresponding detection accuracy. As depicted in Figure 7(b), with an increase in $\gamma$, the number of sample points decreases, facilitating a proportional acceleration in clustering speed. However, the accuracy benefits are not linear with higher $\gamma$ values. When $\gamma = 0.7$, an excessive number of points are disregarded, leading to a significant decline in detection accuracy. Therefore, we select $\gamma = 0.5$ as the threshold parameter.

**The selection of base detector.** In Table 6 , we present a comparison of GFL with several leading base detectors, where GFL achieves superior performance. As indicated by AP_S, both GFL and CenterNet are effective for small object detection, but GFL exhibits more balanced accuracy across various object scales. In contrast to FCOS and YOLOv8, GFL outperforms both in AP and AP_S. Therefore, we choose GFL as the base detector of AD-Det.

**The design of ASOE module.** In our ASOE module, we empirically select the lowest layer of the feature maps, namely P3, as our primary focus [**?**]. In Figure 8 [Figure 8: see original paper], we visualize the regions of interest selected by ASOE using different feature maps. The visualization shows that the P3 layer focuses more on small object regions, while higher layers such as P4 and P5 gradually focus on larger objects. To substantiate its efficacy, we conduct a straightforward comparison with the higher adjacent layer P4 as an alternative focus. We summarize the results in Table 7 . Experimental results indicate that the P3 layer yields the highest performance, while clustering from higher-level layers such as P4 would lead to performance degradation. Note that AP and AP_S perform a sharp decrease from P3 to P4, highlighting our approach' s effectiveness in detecting small objects within higher-resolution layers.

**The design of DCC module.** DCC comprises two major components: diversity augmentation (DA) and dynamic search (DS). Based on ASOE, we verify these two components step by step. As shown in Table 8 , when DA is applied to tail-class instances, the AP_S and AP_M are increased by 0.4% and 0.7%, respectively. When further combined with DS, the AP is increased to 35.9%, which proves its effectiveness.

## 5. Conclusion

In this paper, we propose AD-Det, a novel object detection framework for UAV images, which addresses the challenges of scale variations and class imbalance. AD-Det consists of two key components: Adaptive Small Object Enhancement (ASOE) and Dynamic Class-balanced Copy-paste (DCC). ASOE employs a high-resolution feature map to cluster regions containing small objects, followed by processing them for fine-grained detection. DCC performs reasonable object re-sampling by dynamically pasting tail classes around clustering centers obtained by ASOE. We conducted extensive experiments on VisDrone and UAVDT, and demonstrated that our approach significantly outperforms existing competitive alternatives.

Although our proposed approach achieves satisfactory results on UAV image object detection, several limitations remain to be addressed in our future endeavors. (1) ASOE achieves precise detection of small objects by leveraging global location cues, but the usage of global features is still limited. ASOE ignores the feature interaction between the global image and the local regions, which may help to better understand the scene' s semantic knowledge. (2) DCC conducts uniform sampling copy-paste, but there is room for improvement by considering instance-specific metrics and complex relationships between classes for more effective hard example mining. For the future, we plan to explore more effective ways to handle UAV images under complex scenarios, including occlusion and background clutter issues.

**Author Contributions:** Conceptualization, Z.L. and S.L.; methodology, Z.L.; software, D.P., Y.W.; validation, Z.L., S.L. and D.P.; formal analysis, Z.L and S.L.; investigation, Z.L. and Y.W.; resources, Z.L.; data curation, Z.L.; writing—original draft preparation, Z.L. and S.L.; writing—review and editing, S.L. and W.L.; visualization, Z.L.; supervision, S.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Natural Science Foundation of Fujian Province of China, Youth project (No.2023J05117); Education and Scientific Research Project for Middle-aged and Young Teachers in Fujian Province (No.JAT220017); Natural Science Foundation of Jiangxi Province (No.20232BAB203057).

**Data Availability Statement:** The datasets used in this study are all open source: VisDrone is available at https://github.com/VisDrone/VisDrone-Dataset (accessed on 17 January 2025), and UAVDT can be accessed at https://sites.google.com/view/grli-uavdt (accessed on 17 January 2025).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Osco, L.P.; Junior, J.M.; Ramos, A.P.M.; de Castro Jorge, L.A.; Fatholahi, S.N.; de Andrade Silva, J.; Matsubara, E.T.; Pistori, H.;

Gonçalves, W.N.; Li, J. A review on deep learning in UAV remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* 2021, 102, 102456. https://doi.org/10.1016/j.jag.2021.102456.

2. Zuo, Z.; Liu, C.; Han, Q.L.; Song, J. Unmanned aerial vehicles: Control methods and future challenges. *IEEE/CAA J. Autom. Sin.* 2022, 9, 601–614. https://doi.org/10.1109/JAS.2022.105410.

3. Kurunathan, H.; Huang, H.; Li, K.; Ni, W.; Hossain, E. Machine Learning-Aided Operations and Communications of Unmanned Aerial Vehicles: A Contemporary Survey. *IEEE Commun. Surv. Tutorials* 2024, 26, 496–533. https://doi.org/10.1109/COMST.2023.3312221.

4. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23-28 June 2014.

5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 7-12 December 2015.

6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27-30 June 2016.

7. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 17-24 June 2023.

8. Gao, Z.; Wang, L.; Han, B.; Guo, S. Adamixer: A fast-converging query-based object detector. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18-24 June 2022.

9. Li, Y.; Wang, S. R(Det)2: Randomized Decision Routing for Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18-24 June 2022.

10. Jia, D.; Yuan, Y.; He, H.; Wu, X.; Yu, H.; Lin, W.; Sun, L.; Zhang, C.; Hu, H. DETRs With Hybrid Matching. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 17-24 June 2023.

11. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*, Zurich, Switzerland, 6-12 September 2014.

12. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021, 44, 7380–7399. https://doi.org/10.1109/TPAMI.2021.3119563.

13. Gao, T.; Niu, Q.; Zhang, J.; Chen, T.; Mei, S.; Jubair, A. Global to local: A scale-aware network for remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* 2023, 61, 1–14. https://doi.org/10.1109/TGRS.2023.3245500.

14. Chen, Z.; Ji, H.; Zhang, Y.; Zhu, Z.; Li, Y. High-resolution feature pyramid network for small object detection on drone view. *IEEE Trans. Circuits Syst. Video Technol.* 2023, 34, 475–489. https://doi.org/10.1109/TCSVT.2023.3304145.

15. Biswas, D.; Tešić, J. Small object difficulty (sod) modeling for objects detection in satellite images. *Proceedings of the International Conference on Computational Intelligence and Communication Networks*, Al-Khobar, Saudi Arabia, 4-6 December 2022.

16. Biswas, D.; Tešić, J. Unsupervised domain adaptation with debiased contrastive learning and support-set guided pseudo labeling for remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2024, 17, 3197–3210. https://doi.org/10.1109/JSTARS.2024.3349541.

17. Li, C.; Yang, T.; Zhu, S.; Chen, C.; Guan, S. Density Map Guided Object Detection in Aerial Images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, 14-19 June 2020.

18. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 20-25 June 2021.

19. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 22-29 October 2017.

20. Li, B.; Liu, Y.; Wang, X. Gradient Harmonized Single-Stage Detector. *Proceedings of the AAAI Conference on Artificial Intelligence* 2019, 33, 8577–8584. https://doi.org/10.1609/aaai.v33i01.33018577.

21. Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 2019.

22. Zhang, X.; Izquierdo, E.; Chandramouli, K. Dense and Small Object Detection in UAV Vision Based on Cascade Network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Seoul, Korea (South), 27-28 October 2019.

23. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In *Proceedings of the Advances in Neural Information Processing Systems*, Virtual, 6-12 December 2020.

24. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021, pp. 1–1. https://doi.org/10.1109/TPAMI.2021.3117983.

25. Gupta, A.; Dollar, P.; Girshick, R. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 15-20 June 2019.

26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 2017, 60, 84–90. https://doi.org/10.1145/3065386.

27. Girshick, R. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 7-13 December 2015.

28. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 22-29 October 2017.

29. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21-26 July 2017.

30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In *Proceedings of the European Conference on Computer Vision*, Amsterdam, The Netherlands, 11–14 October 2016.

31. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 8-14 September 2018.

32. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea (South), 27 October–2 November 2019.

33. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, Glasgow, UK, 23–28 August 2020.

34. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In *Proceedings of the*

*IEEE International Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18-24 June 2022.

35. Zhu, Y.; Zhou, Q.; Liu, N.; Xu, Z.; Ou, Z.; Mou, X.; Tang, J. ScaleKD: Distilling Scale-Aware Knowledge in Small Object Detector. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 17-24 June 2023.

36. Han, W.; Li, J.; Wang, S.; Wang, Y.; Yan, J.; Fan, R.; Zhang, X.; Wang, L. A context-scale-aware detector and a new benchmark for remote sensing small weak object detection in unmanned aerial vehicle images. *Int. J. Appl. Earth Obs. Geoinf.* 2022, 112, 102966. https://doi.org/10.1016/j.jag.2022.102966.

37. Cao, B.; Yao, H.; Zhu, P.; Hu, Q. Visible and Clear: Finding Tiny Objects in Difference Map. In *Proceedings of the European Conference on Computer Vision*, Milan, Italy, 29 September–4 October, 2024.

38. Huang, Y.X.; Liu, H.I.; Shuai, H.H.; Cheng, W.H. DQ-DETR: DETR with Dynamic Query for Tiny Object Detection. In *Proceedings of the European Conference on Computer Vision*, Milan, Italy, 29 September–4 October, 2024.

39. Yin, N.; Liu, C.; Tian, R.; Qian, X. SDPDet: Learning Scale-Separated Dynamic Proposals for End-to-End Drone-View Detection. *IEEE Trans. Multimedia* 2024, 26, 7812–7822. https://doi.org/10.1109/TMM.2024.3375398.

40. Zhang, C.; Liu, T.; Xiao, J.; Lam, K.M.; Wang, Q. Boosting Object Detectors via Strong-Classification Weak-Localization Pretraining in Remote Sensing Imagery. *IEEE Trans. Instrum. Meas.* 2023, 72, 1–20. https://doi.org/10.1109/TIM.2023.3315392.

41. Zhang, C.; Xiao, J.; Yang, C.; Zhou, J.; Lam, K.M.; Wang, Q. Integrally Mixing Pyramid Representations for Anchor-Free Object Detection in Aerial Imagery. *IEEE Geosci. Remote Sens. Lett.* 2024, 21, 1–5. https://doi.org/10.1109/LGRS.2024.3404481.

42. Duan, C.; Wei, Z.; Zhang, C.; Qu, S.; Wang, H. Coarse-grained density map guided object detection in aerial images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Montreal, BC, Canada, 11-17 October 2021.

43. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered object detection in aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea (South), 27 October–2 November 2019.

44. Deng, S.; Li, S.; Xie, K.; Song, W.; Liao, X.; Hao, A.; Qin, H. A Global-Local Self-Adaptive Network for Drone-View Object Detection. *IEEE Trans. Image Process.* 2021, 30, 1556–1569. https://doi.org/10.1109/TIP.2020.3044929.

45. Meethal, A.; Granger, E.; Pedersoli, M. Cascaded Zoom-In Detector for High Resolution Aerial Images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, Vancouver, BC, Canada, 17-24 June 2023.

46. Liu, C.; Gao, G.; Huang, Z.; Hu, Z.; Liu, Q.; Wang, Y. YOLC: You Only Look Clusters for Tiny Object Detection in Aerial Images. *IEEE Trans. Intell. Transp. Syst.* 2024, 25, 13863–13875. https://doi.org/10.1109/TITS.2024.3386928.

47. Zhang, C.; Lam, K.M.; Liu, T.; Chan, Y.L.; Wang, Q. Structured Adversarial Self-Supervised Learning for Robust Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 2024, 62, 1–20. https://doi.org/10.1109/TGRS.2024.3375398.

48. Yu, W.; Yang, T.; Chen, C. Towards Resolving the Challenge of Long-Tail Distribution in UAV Images for Object Detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, 3-8 January 2021.

49. Wang, T.; Li, Y.; Kang, B.; Li, J.; Liew, J.; Tang, S.; Hoi, S.; Feng, J. The devil is in classification: A simple framework for long-tail instance segmentation. In *Proceedings of the European Conference on Computer Vision*, Glasgow, UK, 23–28 August 2020.

50. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 15-20 June 2019.

51. Li, Y.; Wang, T.; Kang, B.; Tang, S.; Wang, C.; Li, J.; Feng, J. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 13-19 June 2020.

52. Zhang, S.; Chen, C.; Peng, S. Reconciling Object-Level and Global-Level Objectives for Long-Tail Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, Paris, France, 1-6 October 2023.

53. Hyun Cho, J.; Krähenbühl, P. Long-tail detection with effective class-margins. In *Proceedings of the European Conference on Computer Vision*, Tel Aviv, Israel, 23–27 October 2022.

54. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21-26 July 2017.

55. Bundy, A.; Wallen, L. Breadth-first search. *Catalogue of Artificial Intelligence Tools* 1984, 1, 13–13. https://doi.org/10.1007/978-3-642-96868-

6_{25}.

56. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 8-14 September 2018.

57. Chen, K.e.a. MMDetection: Open MMLab Detection Toolbox and Benchmark. arXiv preprint arXiv:1906.07155 2019.

58. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 2019, 32, 8026–8037.

59. Du, B.; Huang, Y.; Chen, J.; Huang, D. Adaptive Sparse Convolutional Networks With Global Context Enhancement for Faster Object Detection on Drone Images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 17-24 June 2023.

60. Wei, Z.; Duan, C.; Song, X.; Tian, Y.; Wang, H. AMRNet: Chips augmentation in aerial images object detection. arXiv preprint arXiv:2009.07168 2020.

61. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. arXiv preprint arXiv:1904.07850 2019.