remote sensing

MDPI

Article

SMFF-YOLO: A Scale-Adaptive YOLO Algorithm with Multi-Level Feature Fusion for Object Detection in UAV Scenes

Yuming Wang [1,2], Hua Zou [1,*], Ming Yin [2] and Xining Zhang [1]

[1] School of Computer Science, Wuhan University, Wuhan 430072, China; 2115053012@mail.wtu.edu.cn (Y.W.); zhangxn@whu.edu.cn (X.Z.)
[2] School of Electronic and Electrical Engineering, Wuhan Textile University, Wuhan 430077, China; 2015053003@mail.wtu.edu.cn
* Correspondence: zouhua@whu.edu.cn

Abstract: Object detection in images captured by unmanned aerial vehicles (UAVs) holds great potential in various domains, including civilian applications, urban planning, and disaster response. However, it faces several challenges, such as multi-scale variations, dense scenes, complex backgrounds, and tiny-sized objects. In this paper, we present a novel scale-adaptive YOLO framework called SMFF-YOLO, which addresses these challenges through a multi-level feature fusion approach. To improve the detection accuracy of small objects, our framework incorporates the ELAN-SW object detection prediction head. This newly designed head effectively utilizes both global contextual information and local features, enhancing the detection accuracy of tiny objects. Additionally, the proposed bidirectional feature fusion pyramid (BFFP) module tackles the issue of scale variations in object sizes by aggregating multi-scale features. To handle complex backgrounds, we introduce the adaptive atrous spatial pyramid pooling (AASPP) module, which enables adaptive feature fusion and alleviates the negative impact of cluttered scenes. Moreover, we adopt the Wise-IoU(WIoU) bounding box regression loss to enhance the competitiveness of different quality anchor boxes, which offers the framework a more informed gradient allocation strategy. We validate the effectiveness of SMFF-YOLO using the VisDrone and UAVDT datasets. Experimental results demonstrate that our model achieves higher detection accuracy, with AP50 reaching 54.3% for VisDrone and 42.4% for UAVDT datasets. Visual comparative experiments with other YOLO-based methods further illustrate the robustness and adaptability of our approach.

Keywords: object detection; unmanned aerial vehicles; tiny objects; complex scenarios; multi-level feature information fusion

Citation: Wang, Y.; Zou, H.; Yin, M.; Zhang, X. SMFF-YOLO: A Scale-Adaptive YOLO Algorithm with Multi-Level Feature Fusion for Object Detection in UAV Scenes. *Remote Sens.* **2023**, *15*, 4580. https://doi.org/10.3390/rs15184580

Academic Editors: Hossein M. Rizeei, Qi Zhao, Guangliang Cheng and Paolo Tripicchio

Received: 29 August 2023
Revised: 13 September 2023
Accepted: 14 September 2023
Published: 18 September 2023

1. Introduction

Object detection, a critical task in computer vision, involves identifying specific objects and determining their positions. It provides an efficient and accurate method for object identification in images or videos. This task plays a significant role and has broad applications in the field of artificial intelligence. The emergence of unmanned aerial vehicles (UAVs) has revolutionized various domains, including urban monitoring, traffic control, and disaster response. Due to their unique advantages and wide range of application scenarios, UAVs have made object detection a pivotal area of research in the field [1,2].

The significant advancement of object detection can be attributed to the emergence of deep learning [3–8]. These methods have demonstrated impressive accuracy on large-scale object detection datasets consisting of natural images, such as MS COCO [9] and PASCAL VOC [10]. Nevertheless, the performance of existing methods designed for natural images in object detection tasks falls short when applied to UAV-captured images. This is mainly due to the significant differences between the two types of images.

There are four primary factors that contribute to the challenges of object detection in UAV-captured images. Firstly, UAV-captured images have different viewpoints compared

to natural images taken from the ground or conventional perspectives. This introduces diversity in the scale, angle, and shape of objects in UAV images, which makes object detection more challenging. Secondly, UAV-captured images exhibit significant variations in object scales. These images contain both small objects, such as pedestrians and bicycles, as well as large objects like buildings and terrain. This scale variation within the images adds complexity to the object detection task. Thirdly, the backgrounds in UAV-captured images can be complex, which presents environmental factors that further complicate object detection tasks. For example, objects within the images may have textures and colors similar to the background, making them harder to distinguish. Fourthly, UAV-captured images often suffer from low resolution due to operation altitude or device limitations. This low resolution leads to the loss of object details and inadequate information, hindering accurate object detection. Furthermore, factors like wind or clouds may affect the UAV-captured images, which results in image blurring or the presence of noise. These factors collectively contribute to the difficulties faced in object detection in UAV-captured images.

Existing methods for object detection in UAV-captured images are typically designed for specific scenarios, such as ship or vehicle detection. These methods are well-suited for cases where object sizes are similar and backgrounds are uniform. Consequently, these limitations result in reduced detection performance for multi-class objects and an increased risk of missing detections when handling objects with significant scale variations. Moreover, object detection in UAV-captured images often involves a considerable number of small objects. These tiny objects exhibit small sizes and limited details, diminishing their distinctiveness in the image and making them susceptible to confusion with the background or other objects. To address these challenges, researchers have devoted efforts to enhance the performance of tiny object detectors by improving feature representation and optimizing data augmentation techniques [11–14]. Despite the progress made by these methods in improving object detection performance, they still possess certain limitations. Notably, these methods exhibit insufficient generalization capability in the context of multi-class object detection, especially when it involves small objects [15–17]. Additionally, complex environmental factors can lead to erroneous detection results, and existing models may not extract sufficient feature information for small objects. In scenarios with complex, blurred, and contaminated backgrounds, a significant amount of information loss occurs, which poses challenges for detecting these objects.

To overcome the challenges of information loss, adaptability to different object sizes, and complex backgrounds in small object detection, this paper proposes an object detector suitable for UAV-captured images based on the YOLO framework. Specifically, to improve the detection accuracy of small objects, we introduce an additional tiny object prediction head. Moreover, we replace the conventional convolution-based prediction heads with novel prediction heads that combine convolution and Swin Transformer [18], enhancing feature representation. Next, we employ a bidirectional feature fusion approach to aggregate feature information of different resolutions in the backbone network, significantly improving the importance of low-level information in the feature maps. Additionally, we utilize the ELAN module [19] as a fundamental module in the YOLO framework, enabling the network to learn more features and exhibit stronger robustness. Furthermore, we design the adaptive atrous spatial pyramid pooling (AASPP) module to alleviate the influence of complex backgrounds in UAV-captured images and enhance the detection accuracy of multi-scale and multi-class objects. Finally, we redesign the loss function based on the characteristics of the dataset and incorporate a dynamic non-monotonic focal mechanism, which results in better performance.

The main contributions of this paper can be summarized as follows:

1. We present SMFF-YOLO, a novel one-stage detection framework for accurate object detection in UAV images. It introduces a novel type of prediction head that fuses convolution with Swin Transformer, enhancing feature representation by combining global and local information. Furthermore, a specialized prediction head is added to detect tiny objects effectively.

2. To improve object detection across scales, we introduce the adaptive atrous spatial pyramid pooling (AASPP) module. This facilitates cross-scale feature fusion and employs mixed attention mechanisms for enhanced feature information. We also propose the bidirectional feature fusion pyramid (BFFP) model to enhance multi-scale fusion via bidirectional information flow.
3. For better regression anchor accuracy in SMFF-YOLO, we adopt Wise-IoU (WIoU) as the bounding box regression loss. This serves to balance anchor box competitiveness and address gradient issues from low-quality samples, which enhances the model's overall performance.
4. We evaluate the performance of SMFF-YOLO by comparing it with several state-of-the-art detection models on the VisDrone and UAVDT datasets. The experimental results clearly show that our proposed method excels in detecting objects in challenging scenarios characterized by substantial variations in object scales.

2. Related Work

2.1. Traditional Object Detection Methods and Deep Learning-Based Object Detection Methods

The development of object detection methods can be categorized into two main categories: traditional object detection and deep learning-based object detection. Traditional methods typically involve three components: object localization, feature extraction, and feature classification. However, traditional object localization methods have limitations. Since objects can appear at arbitrary locations in an image with uncertain sizes and aspect ratios, the initial approach involves exhaustively sliding windows over the entire image at different scales and aspect ratios. While this exhaustive strategy covers all possible object locations, it has drawbacks, including high time complexity, excessive redundant windows, and adverse effects on subsequent feature extraction and classification speed and performance. Additionally, traditional object detection methods rely on manually designed features, which may not fully capture the complexity of objects. Consequently, traditional methods exhibit relatively poor performance and robustness in complex scenarios.

Deep learning-based object detection methods have emerged as the dominant approach in the field. They can be classified into two categories. The first category is a two-stage detection method based on region proposals. An early example is R-CNN [20], which extracts region proposals and utilizes convolutional neural networks (such as AlexNet [21] and VGG [22]) for object classification and bounding box regression. Fast R-CNN [3] improves detection speed by introducing the ROI (region of interest) pooling layer. Faster R-CNN [3] further enhances the speed and accuracy of detection by introducing the region proposal network, which unifies region proposal extraction and object detection. Mask R-CNN [23] extends Faster R-CNN to support instance segmentation, enabling precise pixel-level masks for objects. Cascade R-CNN [24] adopts a cascading approach that trains and filters multiple detectors in a progressive manner, gradually improving the detection accuracy. Libra R-CNN [25] addresses class imbalance in object detection by introducing balancing strategies and loss functions to mitigate the performance degradation caused by imbalanced classes.

The one-stage approach is another category of methods that directly regress the object size, location, and class from the input image using the network. One of the most widely recognized one-stage detection methods is the YOLO series of frameworks [19,26,27]. These frameworks treat object detection as a regression problem and predict bounding boxes and class information for each grid in the image. SSD [5] achieves multi-scale object detection by predicting targets at various feature map levels. RetinaNet [6] utilizes a feature pyramid network to extract features and addresses class imbalance in object detection. To improve the detection performance of small objects, it introduces focal loss. CornerNet [28] determines the position and size of objects by detecting their keypoints using convolutional neural networks for keypoint prediction.

2.2. Deep Learning-Based Object Detection Methods for UAV-Captured Images

In recent years, the rapid advancement of UAV technology has led to the capturing of images with unique characteristics and challenges. These challenges include variations in different scenes, complex backgrounds, scale variations of objects, and occlusion of targets. These factors pose significant challenges for traditional object detection methods to achieve satisfactory performance when applied to UAV-captured images. Fortunately, the introduction of large-scale datasets specifically designed for UAV-captured images has facilitated significant breakthroughs in object detection by utilizing deep learning-based approaches [29,30].

Convolutional neural networks (CNNs) have significantly advanced object detection in the field by leveraging their automatic feature extraction capabilities. In the domain of UAV-captured images, deep learning has been increasingly integrated into object detection methodologies. Wu et al. [31] improved the spatial pyramid pooling structure by incorporating additional pooling layers and cascading multiple groups of pooling layers, thereby enhancing the learning capacity of the network. Chen et al. [32] proposed a classification-oriented super-resolution generative adversarial network to augment data and detect small vehicles in UAV images. ClusDet [33] introduced an end-to-end framework that unified object clustering and detection, achieving high runtime efficiency and effectively improving the detection accuracy of small objects. AdNet [34] introduced a multi-scale adversarial network that aligns features across different viewpoints, lighting conditions, weather, and backgrounds, which enhances its adaptability. LMSD-YOLO [35] proposed a lightweight SAR ship detection model composed of convolutional blocks using depth-wise separable convolutions, reducing computational demands while accelerating model convergence and improving detection accuracy. BIFA-YOLO [36] designed a detection head with angle classification capability, which enables accurate capture of angle information in ship detection for cases involving boundary continuity and complex parameter regression.

Unlike traditional CNNs, transformers leverage self-attention mechanisms, which allow them to capture global contextual information beyond local features. This ability makes transformers more effective in handling long-range dependencies and global relationships. In the domain of object detection in UAV-captured images, the introduction of the transformer brings several advantages. TPH-YOLOv5 [37] effectively integrates YOLOv5 and transformers, with transformer prediction heads that significantly improve performance in images with significant variations in object sizes. TPH-YOLOv5++ [38] designs a cross-layer asymmetric transformer to replace additional prediction heads, which results in a notable improvement in detection speed while retaining most of the knowledge from the additional prediction heads. VIT-YOLO [39] designs an improved backbone to preserve sufficient global contextual information and extracts more diverse features using multi-head self-attention for object detection. The transformer structure is particularly effective in utilizing global contextual information, which is crucial for understanding the relationship between objects and their surroundings. Additionally, transformers possess strong feature representation abilities, which allows them to capture the intricate details and complexity of objects, thus enhancing detection accuracy. Moreover, the self-attention mechanism of the transformer enables adaptive handling of objects at different scales, which makes the algorithm more robust. In summary, the introduction of transformers introduces new possibilities to the field of UAV applications. In our design, the combination of Swin Transformer and CNN further enhances object detection performance in complex environments.

3. Proposed Method

This section provides an overview of the method proposed in this paper, outlining its overall structure and discussing several specific improvement measures. These measures include the replacement of the original prediction heads with a new prediction head consisting of Swin Transformer and CNN. Additionally, the paper introduces a tiny object

prediction head, a novel spatial pyramid pooling module, a bidirectional feature fusion pyramid module, and an improved loss function.

The architecture of our proposed SMFF-YOLO is depicted in Figure 1, encompassing three main components: backbone, neck, and prediction heads. The backbone network consists of ELAN modules, designed based on gradient path and showcasing strong learning capabilities. Additionally, we introduce the AASPP module, positioned at the end of the backbone, to further enhance feature extraction. In the neck component, we draw inspiration from EfficientDet [40] and devise the BFFP module, efficiently connecting multi-scale features bidirectionally to improve the model's ability to extract features at different scales. Lastly, our model employs four prediction heads to generate outputs. The joint module ELAN-SW, combining Swin Transformer and ELAN, forms the core component of the prediction heads. This integration enables the model to build long-range dependencies and capture global context information in the input image. Consequently, the model can better comprehend the semantics and spatial relationships of the objects.

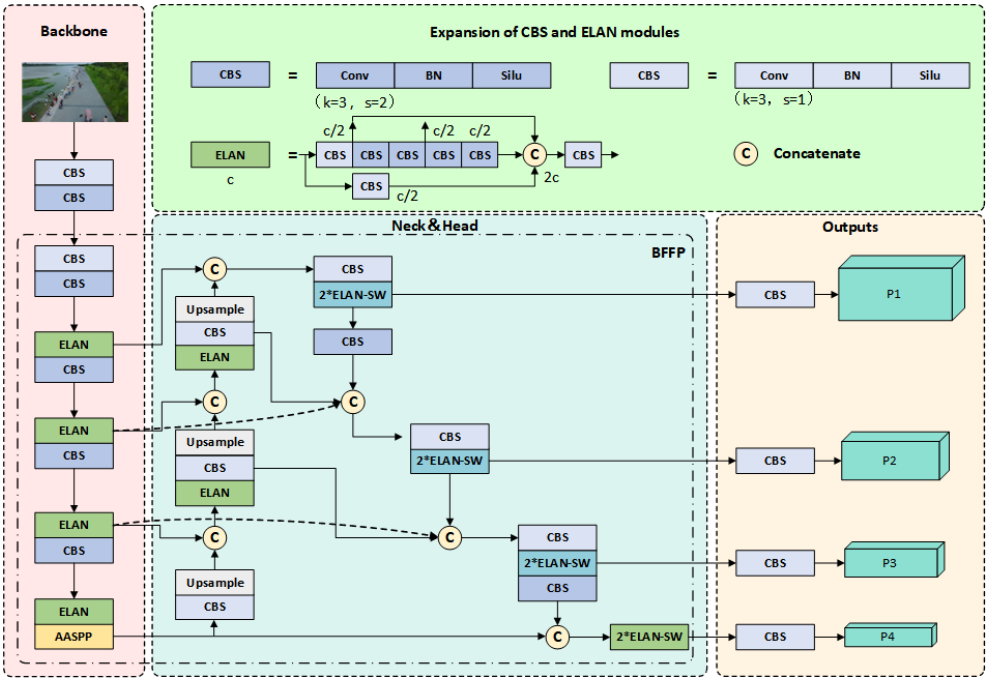

Figure 1. Overall architecture of the proposed SMFF-YOLO.

3.1. Additional Tiny Object and Swin Transform Prediction Heads

In UAV-captured images, objects exhibit a wide range of scale variations, often consisting of many tiny-sized targets. Detecting these tiny objects is challenging due to the limited availability of spatial context information and low visual prominence. Typically, the YOLO framework utilizes three prediction heads to improve the accuracy of detecting tiny-sized targets. However, it is clearly insufficient for handling all possible detection scenarios. To address this issue, we introduce a dedicated prediction head for tiny objects. This prediction head extracts features from high-resolution feature maps, which makes it more sensitive to tiny objects. Through collaboration with the other three prediction heads, we effectively mitigate the negative impact of significant scale variations on the performance of object detection. Our design strategy allows our model to effectively handle object detection tasks with substantial differences in scale.

Moreover, the transformer model has powerful capabilities in modeling long-range dependencies, which aid in understanding the relationships and contextual information among objects, thus improving the accuracy and robustness of object detection. In UAV-captured images, due to the significant variations in object scales, traditional YOLO frameworks that solely rely on convolutional prediction heads often struggle to effectively detect objects of

various sizes. Therefore, drawing inspiration from DETR [7], we combine transformers with convolutional layers to serve as a novel prediction head in our model. Specifically, we connect and stack ELAN modules and Swin Transformer modules twice to generate the output. The structure of ELAN-SW is illustrated in Figure 2. Swin Transformer is a transformer-based method proposed in recent years. It improves the ability to capture contextual information and spatial information in images by introducing the window multi-head self-attention (W-MHSA) module and the shifted window multi-head self-attention (SW-MHSA) module. The W-MHSA module adopts the window attention mechanism, which divides the input feature map into windows to effectively model local information by calculating attention weights within each window. The SW-MHSA module utilizes the shifted window attention mechanism, which considers the relationship between adjacent windows to effectively capture local information. The Swin Transformer module manifests exceptional prowess in contextual modeling by integrating self-attention mechanisms and cross-window shifting operations. This integration enables the module to effectively capture context information at a global scale, thereby facilitating a deeper understanding of the relationships between objects and their surrounding environment. In addition to the benefits of the Swin Transformer, convolutional layers serve as traditional feature extractors with remarkable capabilities in local perception. They extract features from localized regions through sliding windows, capturing fine details and texture information of objects. By combining convolutional layers and Swin Transformer, we can fully leverage the advantages of both methods. This allows the model to consider both global and local information, resulting in an overall performance improvement.

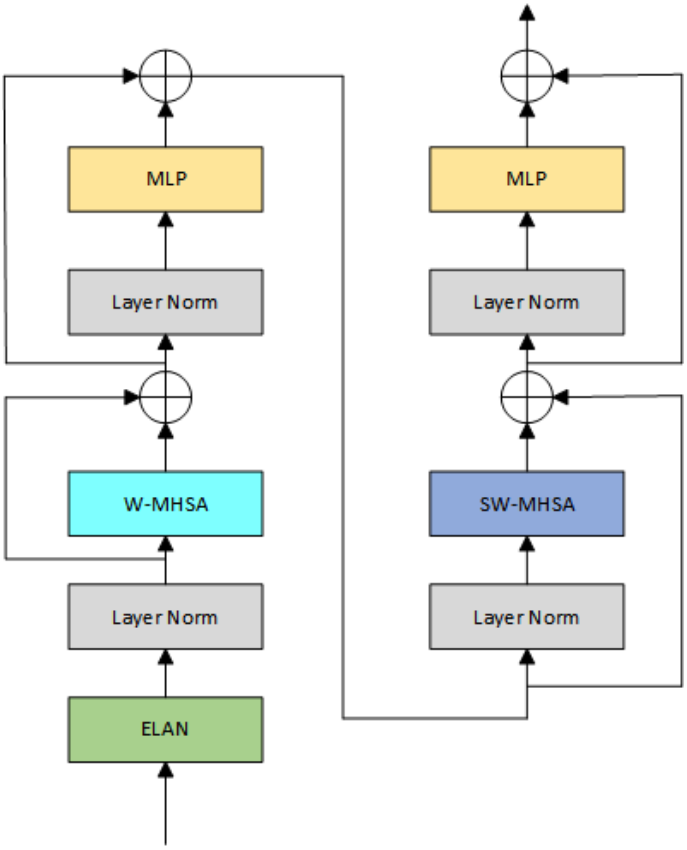

Figure 2. The detailed structure of the ELAN-SW module.

3.2. Adaptive Atrous Spatial Pyramid Pooling Module

In the YOLO framework, it is common to incorporate a pyramid pooling structure at the end of the backbone network to enhance contextual information and thus improve the accuracy and robustness of object detection. However, in recent years, numerous pyramid pooling structures have been proposed, many of which involve multiple pooling operations. These multiple pooling operations can inadvertently result in the loss of scale

information, as objects of varying scales will be compressed into the same length of feature representation after pooling. Consequently, this could lead to inadequate representation of small-scale objects and information compression for large-scale objects.

To overcome this challenge, we propose a more effective solution called the adaptive atrous spatial pyramid pooling (AASPP) module. This module is designed to preserve scale information and enhance the performance of object detection. Figure 3 illustrates the structure of the AASPP module.

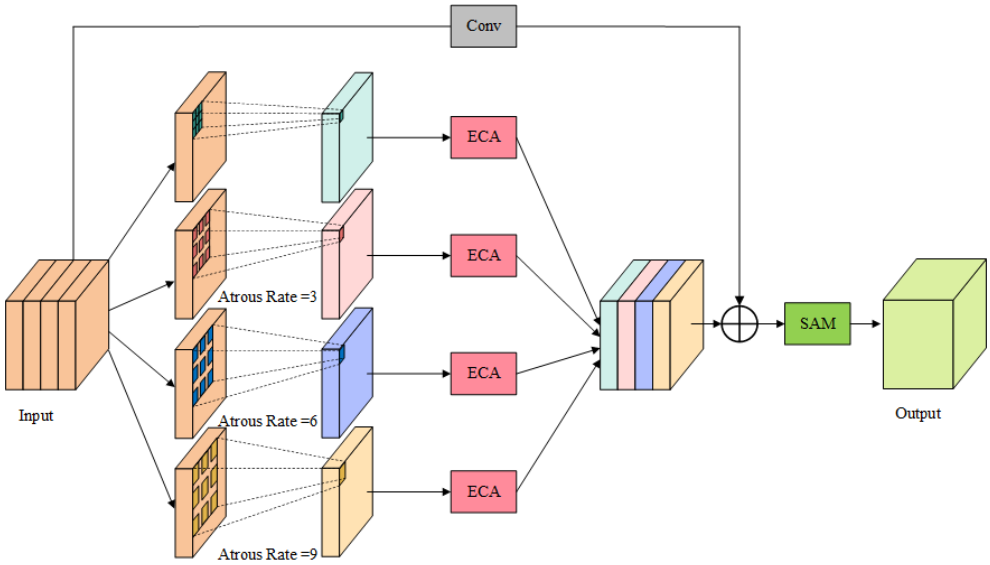

Figure 3. Structure of AASPP module.

The AASPP module is incorporated at the end of the backbone network, similar to other pyramid pooling modules. Firstly, the AASPP module takes the feature map P, which is generated by the deep convolution of the backbone network, as input. It then performs convolution operations using convolutional kernels with different dilation rates, and four feature maps at different scales are generated, denoted as P_i (i = 1, 2, ..., 4). This step enables multi-scale feature extraction, which allows the model to effectively detect objects at various scales and achieve a balance and integration of information. Subsequently, each P_i undergoes further feature enhancement through the use of the enhanced channel attention (ECA) module [41], producing Y_i. The ECA module is illustrated in Figure 4. The output of each P_i processed by the ECA module is denoted as follows:

$$S_i = F_{eca}(P_i, \delta) = \delta(Conv1D(GAP(P_i))) \tag{1}$$

$$Y_i = S_i \cdot P_i, \tag{2}$$

where $GAP(\cdot)$ denotes global average pooling, and $Conv1D(\cdot)$ represents a one-dimensional convolution with a kernel size of k in the channel domain, to simulate local cross-channel interactions. The parameter k determines the coverage range of the interaction. In ECA, the kernel size k is adaptively determined based on the channel dimension C, instead of being manually adjusted, using cross-validation:

$$k = \left| \psi(C) = \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{odd}. \tag{3}$$

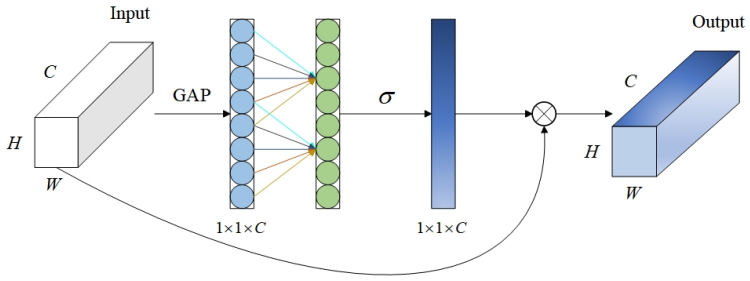

Figure 4. Structure of the ECA module.

Furthermore, all Y_i are concatenated and then subjected to overall feature recalibration using the spatial attention module (SAM) [42]. The spatial attention module optimizes the channel representation of the feature map in an adaptive manner by compressing the channels and performing average pooling and max pooling operations in the channel dimension. Moreover, by calculating the importance weights of each pixel and applying them to every position of the feature map, the spatial attention module helps the network focus on the significant regions in the image. This step further enhances the model's comprehension of the target position and context, resulting in the final output. The spatial attention module is depicted in Figure 5. The final output of the AASPP module, denoted as I, can be defined as follows:

$$I' = f^{1\times1}(P) + f^{3\times3}[Y_1, Y_2, Y_3, Y_4] \tag{4}$$

$$I = \delta(f^{7\times7}[Avg(I'), Max(I')]) \otimes I', \tag{5}$$

where $f^{7\times7}[\cdot]$ represents the convolution operation with a kernel size of 7 and concatenates the elements within it. The term $f^{1\times1}(\cdot)$ represents the convolution operation with a kernel size of 1. The δ stands for the sigmoid activation function. In the AASPP module, each feature map obtained from different atrous convolutions undergoes additional feature enhancement to enhance its discriminative and expressive capabilities. Finally, by integrating all the feature maps that have undergone feature enhancement and incorporating the spatial attention mechanism, the final output feature map is obtained. The inclusion of the AASPP module addresses the problem of scale information loss in traditional pyramid pooling methods, resulting in a more comprehensive and accurate representation of the objects.

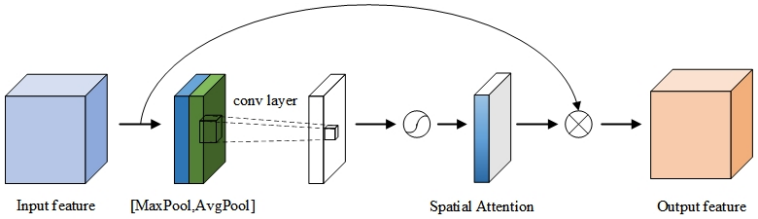

Figure 5. Structure of the spatial attention module.

3.3. Bidirectional Feature Fusion Pyramid

In convolutional networks, deep layers are adept at capturing semantic features, while shallow layers are more responsive to image features. However, this feature representation poses challenges in object detection tasks. On one hand, although deep layers can capture semantic features, their feature maps have low spatial resolutions, which limits the available geometric information for precise object detection. This limitation is especially evident when detecting small objects. Moreover, shallow layers contain more geometric information but lack sufficient semantic features, leading to subpar performance in image classification tasks. To tackle this issue, the feature pyramid network (FPN) [43] was introduced. FPN

incorporates a top-down information propagation mechanism that effectively fuses and represents multi-scale features. PAFPN [44] builds upon FPN by adding a bottom-up pathway, which enables the predicted feature maps to possess both high semantic and positional information. NAS-FPN [45] optimizes the construction of the feature pyramid by autonomously searching for network architectures, thereby enhancing object detection performance. ASFF [46] introduces an adaptive feature fusion mechanism that dynamically allocates weights for feature fusion based on the quality and contribution of each scale's feature map to object detection. Recursive-FPN [47] proposes a recursive feature fusion method for more efficient handling of multi-scale features. To further optimize the multi-scale feature fusion in FPN, BIFPN [40] introduces a bidirectional feature fusion approach. Inspired by the structure of BIFPN and in combination with YOLO, we propose a novel module called bidirectional feature fusion pyramid (BFFP). BFFP incorporates top-down feature propagation from higher-level features and bottom-up feature propagation from lower-level features, achieving a bidirectional flow of information. Figure 6 illustrates the structure of the BFFP module.

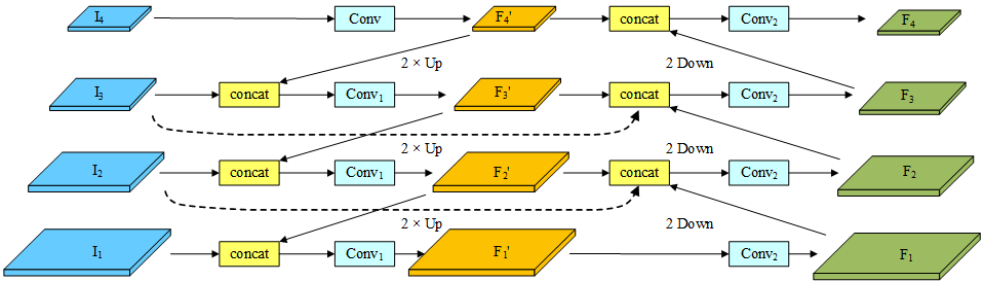

Figure 6. Structure of the BFFP module.

To simplify the description, we will refer to the ELAN module as the $Conv_1$ module and ELAN-SW as the $Conv_2$ module, as depicted in Figure 6. Our proposed BFFP module takes four feature maps extracted by the backbone network as inputs, denoted as $I_1 - I_4$ from bottom to top. Among these, I_1 has the highest feature map resolution, while the subsequent feature maps have their resolutions halved. To enhance the framework's ability to fuse multi-scale features and minimize framework complexity, we chose to incorporate bidirectional skip connections between the middle two layers of the module in our practical experiments. This allows us to effectively capture feature information across different scales while mitigating issues related to excessive parameter increase, model oversizing, gradient vanishing, and feature degradation. The process of fusing multi-scale features in the BFFP module can be described as follows:

$$F_4' = Conv(I_4) \tag{6}$$

$$F_3' = Conv_1[I_3, Up(F_{4'})] \tag{7}$$

$$F_2' = Conv_1[I_2, Up(F_{3'})] \tag{8}$$

$$F_1' = Conv_1[I_1, Up(F_{2'})] \tag{9}$$

$$F_1 = Conv_2(F_1') \tag{10}$$

$$F_2 = Conv_2[Down(F_1), I_2, F_2'] \tag{11}$$

$$F_3 = Conv_2[Down(F_2), I_3, F_3'] \tag{12}$$

$$F_4 = Conv_2[Down(F_3), F_4'], \tag{13}$$

where $F_1' - F_4'$ represent the feature maps obtained from the top-down path, corresponding to levels 1 to 4, while $F_1 - F_4$ represent the output feature maps from the bottom-up path.

The notation *Conv* denotes a 1x1 convolution operation, $Conv_1$ implies that the feature map undergoes the $Conv_1$ module (simplified ELAN module), and $Conv_2$ indicates that the feature map undergoes the $Conv_2$ module (simplified ELAN-SW module). The operation *Up* represents upsampling, and *Down* denotes a convolution operation with a kernel size of 3 and a stride of 2.

3.4. Loss Function

Wise-IoU (WIoU) [48] is a loss function used in object detection tasks to regress boundary boxes. Traditional methods often assume that the majority of examples in the training dataset are of high quality and focus on improving the boundary box regression for these examples. However, real-world object detection datasets, especially those from UAV-captured images like ours, often contain examples of low quality. Overemphasizing boundary box regression for these low-quality examples can have a negative impact on detection performance. In contrast to other boundary box loss functions such as DIoU [49], EIoU [50], CIoU [49], and SIoU [51], WIoU introduces a dynamic non-monotonic focus mechanism. This mechanism allows WIoU to prioritize ordinary-quality boundary boxes, thereby improving the overall performance of the detector. The overall loss of the network can be defined as follows:

$$L = W_{box}L_{box} + W_{cls}L_{cls} + W_{obj}L_{obj} \tag{14}$$

$$L_{box} = L_{WIoU}, \tag{15}$$

where L_{box} represents the bounding box loss, also known as L_{WIoU}, L_{cls} represents the category loss, and L_{obj} represents the confidence loss. W_{box}, W_{cls}, and W_{obj} refer to the weights assigned to the individual losses, respectively. The total loss L is calculated by taking the weighted sum of the three losses.

The bounding box regression model is shown in Figure 7 where, for the anchor box $B = [x, y, w, h]$, the values correspond to the center coordinates and size of the bounding box. Similarly, $B_{gt} = [x_{gt}, y_{gt}, w_{gt}, h_{gt}]$ describes the properties of the target frame. The L_{WIoU} can then be further interpreted as:

$$L_{WIoU} = rL_{IoU}R_{WIoU} \tag{16}$$

$$L_{IoU} = 1 - IoU \tag{17}$$

$$IoU = 1 - \frac{W_iH_i}{wh + w_{gt}h_{gt} - W_iH_i} \tag{18}$$

$$R_{WIoU} = exp(\frac{(x - x_{gt})^2 + (y - y_{gt}^2)}{(W_g^2 + H_g^2)^*}), \tag{19}$$

where *IoU* represents the degree of overlap between anchor boxes and target boxes in object detection tasks, and R_{WIoU} is a distance attention mechanism. To prevent interference of the distance attention mechanism R_{WIoU} with gradients, we detach W_g, H_g, and the computation graph, denoted by superscript *. In Equation (19), we weaken the influence of geometric factors on the loss, thereby improving the model's generalization ability. At the same time, we limit the range of R_{WIoU}, L_{WIoU} ($R_{WIoU} \in [1, e)$, and $L_{WIoU} \in [0, 1]$), which increases the L_{WIoU} of ordinary quality anchor boxes and decreases the R_{WIoU} of high-quality anchor boxes. Through such adjustments, the loss function will pay more attention to the distance between the center points of anchor boxes and target boxes:

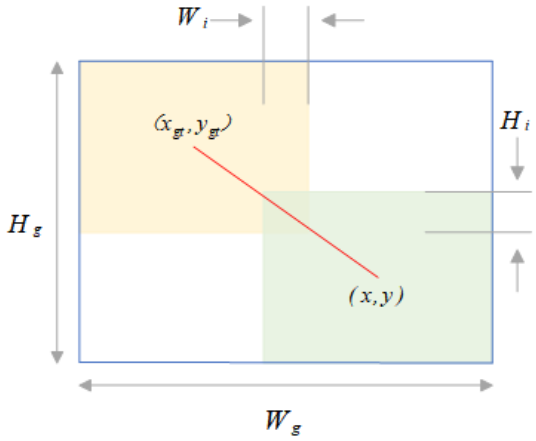

Figure 7. Bounding box regression model.

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{20}$$

$$\beta = \frac{L^*_{IoU}}{\overline{L_{IoU}}}, \tag{21}$$

where β is introduced to define the outlier factor to describe the quality of anchor boxes, and $\overline{L_{IoU}}$ is the sliding average with momentum m. Therefore, the quality threshold for anchor boxes is also dynamic, enabling L_{WIoU} to dynamically allocate gradient gains according to the current situation. Moreover, in order to avoid negative gradients caused by low-quality examples, we utilize β to construct a dynamic non-monotonic focus coefficient r. The dynamic non-monotonic focal mechanism can effectively reduce the competitiveness of high-quality anchor boxes while also reducing the harmful gradients produced by low-quality examples, thereby improving the performance of the model. In this paper, we set α and δ as hyperparameters, with values of 1.9 and 3, respectively, based on our experimental findings and the recommendations of Tong et al. [48]. Furthermore, in our experiments, we set the momentum m to 0.00001.

4. Experiments

The current section commences by providing an overview of the experimental datasets, implementation details, and the evaluation metrics employed. Following that, a thorough analysis is presented, focusing on the datasets' specific characteristics. Lastly, extensive experimental results are provided, showcasing the efficacy and superiority of the proposed method.

4.1. Dataset Introduction

In the research into object detection in UAV-captured images, one of the main and commonly used datasets is the VisDrone dataset [30]. This dataset consists of three parts: the training set, validation set, and test set. The training set contains 6471 images, the validation set contains 548 images, and the test set contains 1610 images. The maximum resolution of the images in the entire dataset is 2000 × 1500. The VisDrone dataset covers multiple object categories, including pedestrians, various types of vehicles, bicycles, motorcycles, and other common classes, totaling 10 classes. We used the VisDrone dataset as the primary research subject for training and evaluating our object detection algorithm. Additionally, we supplemented our research with the UAVDT dataset [29] for ablation experiments. The UAVDT dataset consists of 50 videos, containing a total of 40,376 images. Referring to the TPH-YOLOv5 [38], we used 24,778 images for training and 15,598 images for testing in the UAVDT dataset. All images have a resolution of 1024 × 540. Furthermore, according to the dataset requirements, we grouped all images from the same video into either the training set

or the testing set. Specifically, among the 50 videos, images from 31 videos were placed in the training set, while images from the remaining 19 videos were placed in the testing set. By using these two datasets, we were able to comprehensively cover diverse scenes and target categories, enabling us to evaluate and compare the performance of our object detection algorithm across different datasets.

In order to conduct a more comprehensive investigation into the object detection problem for UAV-captured images, a thorough survey and analysis of the VisDrone dataset and the UAVDT dataset were carried out. Both datasets comprise real-world images captured by UAVs, encompassing diverse urban and rural areas with varied scenes, lighting conditions, and object categories. The research findings revealed common challenges shared by these datasets, which can be categorized into three aspects. Firstly, the prevalence of small objects and occluded targets in the images poses a significant challenge. These objects often have small sizes and can be partially or fully occluded by the surrounding environment, other objects, or obstructions, making accurate detection difficult. This challenge is exemplified in Figure 8a. Secondly, due to variations in the UAV's capturing angles and heights, the targets in the images appear with different scales, angles, and shapes. This diversity adds complexity to the object detection task, necessitating models with good scale adaptability and pose robustness. Figure 8b illustrates these variations. Lastly, the objective weather conditions also affect the quality of UAV-captured images. Differences in lighting, brightness, and contrast introduce variations in image quality. Specific weather conditions like foggy weather and nighttime images present additional challenges, as depicted in Figure 8c. By comprehensively understanding these challenges present in the datasets, it becomes possible to effectively design and improve object detection algorithms, enhancing accuracy, robustness, and adaptability in detecting objects in UAV-captured images.

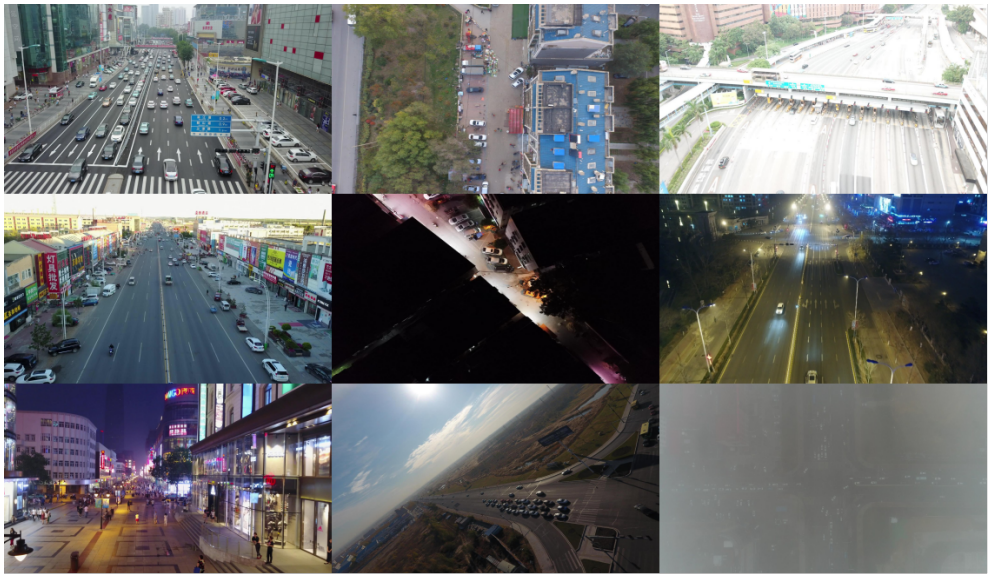

Figure 8. UAV-captured images in the VisDrone dataset and UAVDT dataset. (**a**) The challenge of tiny objects and occluding objects. (**b**) The challenge of UAV imaging perspectives. (**c**) The challenge of objective weather conditions.

4.2. Implementation Details

SMFF-YOLO is built upon the architecture of YOLOv7. Given the distinctive characteristics of UAV-captured images, we made necessary adaptations to the original YOLOv7 framework. For future experiments, the results of the modified YOLOv7 will serve as the baseline. The experiments were conducted using an Intel® Xeon® Silver 4310 CPU operating at 2.10 GHz and NVIDIA RTX3090 graphics cards with 24 GB of video mem-

ory. The experimental software was configured with PyTorch 1.12 and CUDA 11.7. The training phase comprised 200 epochs, utilizing the first 2 epochs for warm-up. The initial learning rate was set to 3.2×10^{-4} and decayed to 0.12 times its value in the final epoch. Parameter optimization was performed using the Adam optimization algorithm [52]. Given the higher resolutions of images in the VisDrone dataset, we configured the input size to be 1536×1536 and utilized a batch size of 4. We conducted post-processing during the testing phase using non-maximum suppression (NMS) with a threshold of 0.6 to derive the final experimental results. Furthermore, for experiments involving the UAVDT dataset, an input size of 1024×1024 was chosen to preserve image information and prevent loss caused by size considerations.

4.3. Evaluation Metrics

In this paper, several evaluation metrics were utilized to assess the performance of our proposed method, including precision (P), recall (R), F1-score ($F1$), average precision at IoU threshold of 0.5 ($AP50$), and mean average precision (AP). Precision measures the ratio of correctly detected objects to all detected results, while recall measures the ratio of correctly detected objects to all ground truth objects. F1-score is a combined measure that takes into account both precision and recall, providing a comprehensive evaluation of model performance. The calculation methods for precision, recall, and F1-score are as follows:

$$P = \frac{TP}{TP + FP} \tag{22}$$

$$R = \frac{TP}{TP + FN} \tag{23}$$

$$F1 = 2\frac{PR}{P + R}, \tag{24}$$

where TP (true positives) represents the number of correctly detected objects, FP (false positives) represents the number of incorrectly detected objects, and FN (false negatives) represents the number of missed objects during the detection process. AP (average precision) is the average precision obtained at different IoU thresholds ranging from 0.5 to 0.95, with an interval of 0.05. The calculation of AP is defined as follows:

$$AP = \int_0^1 P(R)dR. \tag{25}$$

In addition to precision, recall, F1-score, and AP, other metrics such as model parameters and FLOPs (floating-point operations) can also be used as reference points for evaluating a model's performance. Model parameters and FLOPs provide insights into the complexity of the model to some extent.

4.4. Analysis of Results

4.4.1. Effect of Additional Tiny Object and Swin Transform Prediction Heads

To improve the detection of targets at different scales, we incorporated an additional tiny object prediction head. In our drone-based detection task, the targets exhibit significant scale variations. Using a limited number of prediction heads may not effectively capture targets of different scales, particularly tiny ones. To evaluate the performance of the modified YOLOv7 with varying numbers of prediction heads, we conducted experiments. The results of these experiments are presented in Table 1.

Table 1. Comparison of our method with different numbers of prediction heads. The best results are highlighted in bold.

Prediction Head	Size	AP50 (%)	AP (%)
P4	48×48	35.0	18.1
P3	98×98	46.5	27.1
P3, P4	$48 \times 48, 98 \times 98$	47.6	27.8
P2	196×196	49.5	31.1
P2, P3, P4	$48 \times 48, 98 \times 98, 196 \times 196$	51.8	31.7
P1	392×392	36.3	24.5
P1, P2, P3, P4	$48 \times 48, 98 \times 98, 196 \times 196, 392 \times 392$	**52.7**	**32.5**

The results shown in Table 1 indicate that using multiple prediction heads in our experiments has produced positive outcomes. We maintained consistency in the model parameters throughout the experiments, only altering the number of prediction heads. Varying the number of prediction heads allowed us to analyze how different scenarios affected the detection performance of the model. Among the experiments with only one prediction head, the P2 prediction head achieved the highest performance, with an AP50 of 49.5% and an AP of 31.1%. However, when comparing the experiments with a single prediction head to those with multiple prediction heads, the performance of the latter was notably superior to that of the former. In the experiments involving P3 and P4 prediction heads, the P3 prediction head demonstrated the highest performance, with an AP50 of 46.5% and an AP of 27.1%. However, the performance further improved when combining P3 and P4 prediction heads, resulting in an AP50 of 47.6% and an AP of 27.8%. Ultimately, our method achieved the optimum performance by employing 4 prediction heads, with an AP50 of 52.7% and an AP of 32.5%. By fusing multiple detection heads, our method proficiently integrates multi-scale feature information. The fusion of multi-scale features enhances the model's ability to represent and detect objects of varying scales. Due to the potential introduction of excessive parameters and FLOPs, we made the decision not to include an additional prediction head.

We compared the experimental results of the adjusted YOLOv7 with different prediction heads, as shown in Table 2. Our study introduces the ELAN-SW module as a newly developed and innovative prediction head, aiming to improve the model's perception of feature information. By incorporating the ELAN-SW module as a replacement for the traditional convolutional prediction head, the table highlights the superior detection performance achieved by our proposed method. The ELAN-SW prediction head showcases substantial advantages compared to the original convolutional prediction head.

Table 2. Experimental results of the adjusted YOLOv7 with additional tiny object prediction heads and replaced prediction heads. The best results are highlighted in bold.

Method	P (%)	R (%)	F1	AP50 (%)	AP (%)
YOLOv7	58.2	52.3	55.1	51.8	31.7
YOLOv7-4Heads	58.8	53.4	56.0	52.7	32.5
YOLOv7-4Heads-SW	**60.3**	**53.7**	**56.8**	**53.7**	**33.3**

4.4.2. Effect of AASPP Module

Table 3 displays the test results obtained from the YOLOv7-4Heads-SW model, where we replaced the spatial pyramid pooling module of the original model with our AASPP module. The AASPP module is designed with varying numbers of atrous convolutions and dilation rates.

Table 3. Experimental results of AASPP with different parameters. The best results are highlighted in bold.

Experiment	Number of Atrous Convolutions	Dilation Rates	AP50 (%)	AP (%)
I	3	3, 5, 7	53.9	33.2
II	3	3, 6, 9	**54.1**	**33.5**
III	3	4, 8, 12	53.7	33.1
IV	3	5, 10, 15	53.5	32.9
V	3	6, 12, 18	53.8	33.3
VI	3	3, 6, 9, 12	54.0	33.4
VII	3	4, 8, 12, 16	53.9	33.3

Based on the results obtained from experiments II, III, VI, and VII, we have observed that increasing the number of dilated convolutions has a certain impact on the model's performance. However, it appears to have a relatively minor effect on AP and may lead to an increase in the model's parameter count. To investigate this further, we conducted several control experiments with a fixed number of atrous convolutions. Through these experiments, we discovered that the AASPP module with dilation rates of 3, 6, and 9 achieved the best performance for small object detection in UAV-captured images. This module resulted in the highest AP, with AP50 reaching 54.1% and AP reaching 33.5%. These findings provide additional evidence that appropriately utilizing atrous convolutions can significantly enhance the model's performance in this specific task.

Utilizing the YOLOv7-4Heads-SW model as the baseline, our primary focus was to compare various popular spatial pyramid modules. Specifically, the compared modules include SPP [53], SPPF, ASPP [54], SPPCSPC [19], and SPPFCSPC [55].

As shown in Table 4, it is evident that the AASPP module surpasses other methods in terms of detection performance. When integrating the AASPP module into the YOLOv7-4Heads-SW model, we achieve the following metrics: precision (59.5%), recall (54.3%), F1-score (56.8), AP50 (54.1%), and AP (33.5%). When compared to the SPPCSPC module, the AASPP module improves these metrics by 0.2%, 0.7%, 0.3, 0.5%, and 0.4%, respectively. Furthermore, the AASPP module reduces parameter count and complexity in relation to the original YOLOv7 framework utilizing the SPPCSPC module by 5.1 G FLOPs and 9.9 M parameters. Despite a slight increase in parameter count and model complexity compared to SPP, SPPF, and ASPP modules, our method demonstrates a clear advantage in terms of detection performance. This is attributed to the AASPP module's effective utilization of appropriate dilation rates for dilated convolutions and the integration of the mixed attention mechanism. By selecting appropriate dilation rates, the AASPP module can better capture the feature information of objects at different scales within the given image. Simultaneously, the mixed attention mechanism enables the model to concentrate on target regions while effectively reducing background interference. This design strategy empowers the AASPP module to play a crucial role in enhancing object detection performance.

Table 4. Comparison of different spatial pyramid structures (tested on the VisDrone test dataset). The best results are highlighted in bold.

Method	P (%)	R (%)	F1	AP50 (%)	AP (%)	FLOPs (G)	Params (M)
SPP	59.5	53.3	56.2	53.7	33.0	**0.9**	**2.6**
SPPF	59.1	53.7	56.3	53.8	33.1	**0.9**	**2.6**
ASPP	59.5	54.1	56.7	53.9	33.3	7.0	10.2
SPPCSPC	59.7	53.6	56.5	53.6	33.1	21.5	28.3
SPPFCSPC	58.9	53.3	56.0	53.3	32.9	21.5	28.3
AASPP	**59.5**	**54.3**	**56.8**	**54.1**	**33.5**	16.4	18.4

4.4.3. Effect of the BFFP Module

In this section, we compare the detection performance of our proposed model with different improved versions of the YOLOv5 and YOLOv7 models. These improved models incorporate all the enhancements that were previously mentioned, with the exception of the BFFP module. Table 5 presents a comparison of the detection results among the different models. Our SMFF-YOLO model, which includes the BFFP module, achieves the best results in terms of precision, recall, F1-score, AP50, and AP when compared to the four mainstream improved models: YOLOv5m, YOLOv5l, YOLOv5x, and YOLOv7. Specifically, when compared to the improved version of YOLOv5l, our model demonstrates improvements of 6.2%, 4.4%, 5.2%, 4.9%, and 5.1% on these metrics, respectively. These results highlight the effectiveness and feasibility of the strategies proposed in our model.

Table 5. Comparison of different models with all the proposed improvements on the VisDrone dataset. The best results are highlighted in bold.

Method	P (%)	R (%)	F1	AP50 (%)	AP (%)
YOLOv5m-improved	52.3	49.6	50.9	47.1	26.2
YOLOv5l-improved	53.6	50.1	51.8	49.4	28.6
YOLOv5x-improved	55.8	52.7	54.4	51.6	30.3
YOLOv7-improved	59.5	54.3	56.8	54.1	33.5
SMFF-YOLO	**59.8**	**54.5**	**57.0**	**54.3**	**33.7**

4.4.4. Comparison with State-of-the-Art Methods

To evaluate the performance of SMFF-YOLO, we compared its experimental results with those of nine other state-of-the-art methods on the VisDrone dataset.

As shown in Table 6, our SMFF-YOLO achieved an average precision (AP) of 54.3% and an average precision at IoU = 0.5 (AP50) of 33.7%, obtaining the highest values for both metrics. In comparison to the baseline method, SMFF-YOLO exhibited significant improvements in AP (increasing from 52.3% to 54.3%) and AP50 (increasing from 31.0% to 33.7%). The enhanced detection accuracy can be attributed to the inclusion of the ELAN-SW detection head and AASPP module in SMFF-YOLO. These components significantly improve the feature extraction ability for detecting objects of various scales in UAV-captured images, while also reducing the occurrence of false positives. In terms of model complexity, our model has a computational complexity of 257.7 G FLOPs and contains 99.1 M parameters. To capture comprehensive feature information for small objects in complex scenes, we increased the depth of the network and implemented advanced prediction heads, which account for the higher computational complexity of our method. Compared to Grid GDF, our method has a similar model complexity. However, when evaluated using AP50 and AP metrics, our method outperforms Grid GDF by 23.5% and 15.5%, respectively, highlighting its superior accuracy. Experimental comparisons demonstrate that a slight rise in complexity can result in a considerable enhancement in performance.

To further validate our method, we conducted experiments on the UAVDT dataset, as shown in Table 7. Compared to current methods, our SMFF-YOLO achieved new SOTA results in all three metrics, with AP50, AP75, and AP being 42.4%, 33.6%, and 28.4%, respectively. SMFF-YOLO outperformed other advanced methods by at least 3% in performance improvement. In conclusion, our method demonstrates excellent detection accuracy for tiny targets and dense scenes, overcoming challenges in object detection in complex scenarios to a certain extent.

Table 6. Comparison of experimental results with nine other state-of-the-art methods on VisDrone2019-DET-test-dev. The best results are highlighted in bold.

Method	AP50 (%)	AP (%)	FLOPs (G)	Params (M)
Faster R-CNN [3]	31.0	17.2	118.8	41.2
Cascade ADPN [56]	38.7	22.8	547.2	90.8
Cascade-RCNN [24]	38.8	22.6	146.6	69.0
RetinaNet [6]	44.3	22.7	**35.7**	**36.4**
Grid GDF [57]	30.8	18.2	257.6	72.0
SABL	41.2	25.0	145.5	99.6
YOLOv5l	42.4	26.6	107.8	46.2
TPH-YOLOv5++ [38]	52.5	33.5	207.0	-
YOLOv7 [19]	48.5	28.1	104.7	36.9
SMFF-YOLO	**54.3**	**33.7**	257.7	99.1

Table 7. Comparison of experimental results with other advanced methods on the UAVDT dataset. The best results are highlighted in bold.

Method	AP50 (%)	AP75 (%)	AP (%)
ClusDet [33]	26.5	12.5	13.7
Zhang et al. [58]	-	-	17.7
GDFNet [57]	26.1	21.7	15.4
GLSAN [59]	30.5	21.7	19.0
DMNet [60]	24.6	16.3	14.7
DSHNet [61]	30.4	19.7	17.8
CDMNet [62]	35.5	22.4	20.7
SODNet [14]	29.9	18.0	17.1
UFPMP-Net [63]	38.7	28.0	24.6
SMFF-YOLO	**42.4**	**33.6**	**28.4**

4.4.5. Summary of Experimental Results

In this section, we discuss the experimental results. Firstly, as shown in Tables 1 and 2, it is evident that, by replacing the original prediction head of YOLOv7 with our novel designed prediction head and adding an additional prediction head for tiny objects, we achieved a substantial increase in the AP score for tiny objects in UAV-captured images, reaching 53.7%. As indicated in Tables 3 and 4, the utilization of appropriate atrous convolutions and attention mechanisms further enhances the detection performance for UAV-captured images. This results in an AP50 of 54.1%, showcasing the distinct advantages of our method over other pyramid structures. In Table 5, the model attains the highest accuracy when using the BFFP module that we designed. We also conducted comparisons with other advanced methods, as presented in Tables 6 and 7. Regarding model complexity, our method does impose a higher computational burden, but it excels in terms of detection performance. On the VisDrone dataset and UAVDT dataset, we achieved AP scores of 54.3% and 28.4%, respectively. These figures validate the effectiveness of our method.

4.5. Discussion

During the experiment, the SMFF-YOLO framework demonstrates promising potential in detecting occluded objects. To illustrate the detection performance results of our method in the presence of occlusion, we meticulously selected 30 images from the VisDrone dataset that include a large number of occluded objects and performed comparative experiments with methods based on YOLO. We assessed the performance of each method, considering precision (P), recall (R), and F1-score.

Figure 9 presents the visualization results of our method and other YOLO-based methods in occluded scenes. The annotated regions of interest, highlighted by yellow rectangles, are enlarged in these images. In the occlusion scenario, SMFF-YOLO (b) is much higher than the other two methods in terms of the proportion of true positive targets. The

experiments clearly demonstrate that our method outperforms other methods in accurately detecting occluded tiny objects, as shown in Table 8. Promisingly, our method outperformed other methods in terms of precision and F1-score, indicating its higher accuracy and overall performance in detecting occluded objects. It is important to note that achieving high recall in scenarios with occluded targets is a common challenge that any object detection algorithm faces. Occlusion remains a difficult problem to overcome. The advantages of our method in terms of precision and F1-score demonstrate its effectiveness in addressing the detection of occluded objects, offering a robust solution for handling the complexities of real-world scenarios. In future research, we will continue to explore optimizations to further enhance the algorithm, aiming for more comprehensive and accurate object detection results, particularly in challenging scenarios involving occlusions.

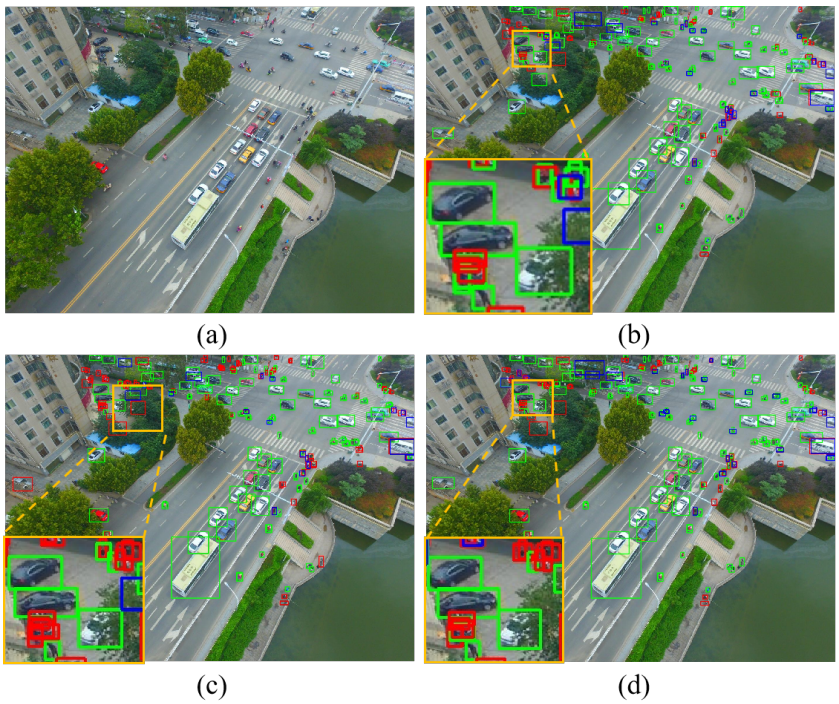

Figure 9. Visual results of different methods based on YOLO in occlusion scenes. (**a**) Original image. (**b**) Results of SMFF-YOLO. (**c**) Results of YOLOv5l. (**d**) Results of YOLOv7. Note that the green bounding boxes represent true positive (TP) targets, the blue bounding boxes represent false positive (FP) targets, and the red bounding boxes represent false negative (FN) targets.

Table 8. Comparative experimental results in scenarios with occluded objects. The best results are highlighted in bold.

Method	P (%)	R (%)	F1
YOLOv5l	57.8	**80.5**	67.9
YOLOv7	61.1	78.7	68.8
SMFF-YOLO	**65.0**	76.5	**70.3**

To comprehensively assess the detection performance of our proposed method, we thoroughly evaluated the effectiveness of SMFF-YOLO using the VisDrone dataset, which consists of real-world scenarios and objects captured by UAVs. Comparative experiments were conducted, with YOLOv5l and YOLOv7 serving as benchmarks. The results of these experiments are visually presented in Figure 10.

Figure 10 illustrates the experimental results obtained from three representative scenes selected from the VisDrone dataset. In these images, specific areas of interest are annotated and enlarged, denoted by yellow rectangles. In Scenario 1 (a), our proposed method, SMFF-YOLO (b), demonstrates superior performance compared to other models in accurately

detecting small objects concealed in the background. Other models often encounter false negatives in such situations. Moving to Scenario 2 (e), we observe that YOLOv5l (g) struggles to effectively detect targets in dense, low-light crowds, resulting in a combination of false positives and false negatives. YOLOv7 (h), while detecting all targets within the specific area, produces a higher number of false positives. Conversely, our proposed method (f) correctly detects all targets present in the specific area. Finally, in Scenario 3 (i), some targets within the specific area are occluded by other objects. Due to the complexity of the background and the small size of the targets, other methods often fail to detect these occluded objects, while our method successfully identifies them. Overall, the experimental results depicted above demonstrate the excellent detection performance of our proposed method on UAV-captured images.

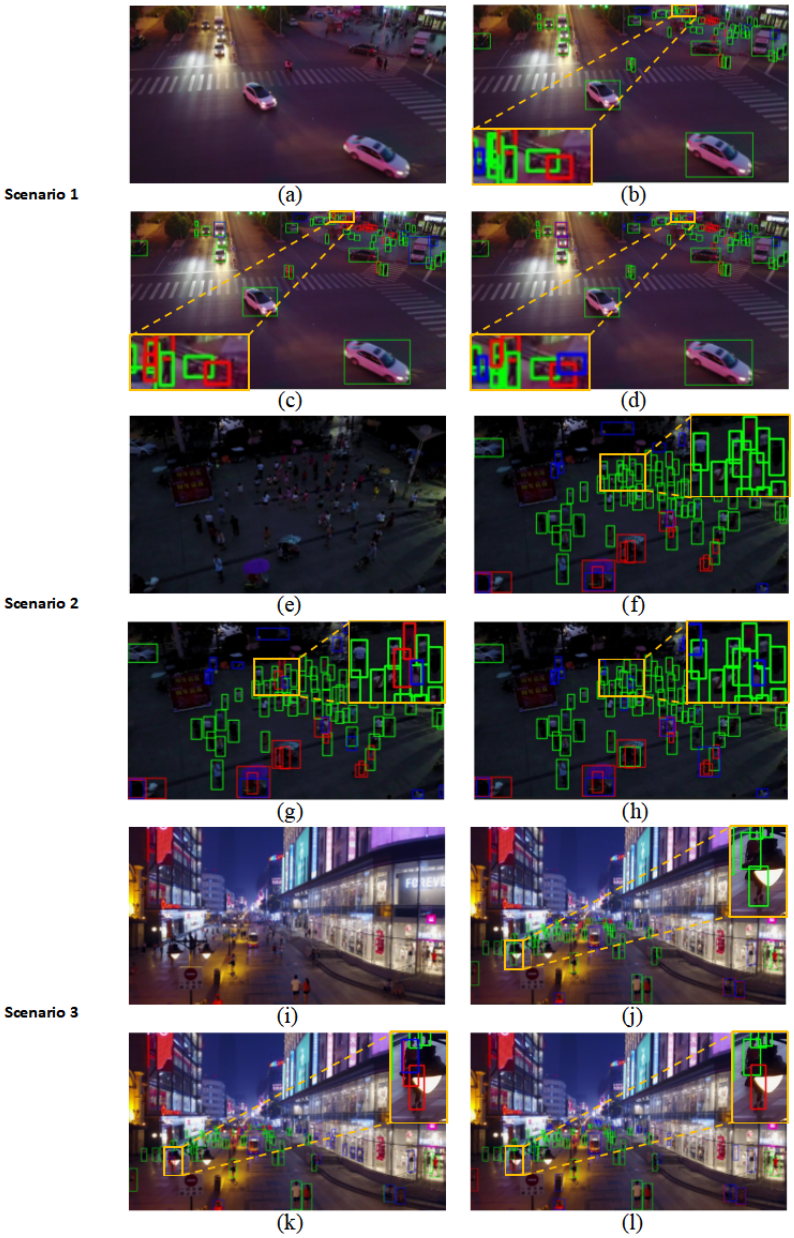

Figure 10. Visual results of different methods based on YOLO on the VisDrone dataset. (**a**,**e**,**i**) Original image. (**b**,**f**,**j**) Results of SMFF-YOLO. (**c**,**g**,**k**) Results of YOLOv5l. (**d**,**h**,**l**) Results of YOLOv7. Note that the green bounding boxes represent true positive (TP) targets, the blue bounding boxes represent false positive (FP) targets, and the red bounding boxes represent false negative (FN) targets.

In addition, we performed a bar chart analysis of the detection results obtained from the three representative scenes depicted in Figure 10. This analysis is visually presented in Figure 11.

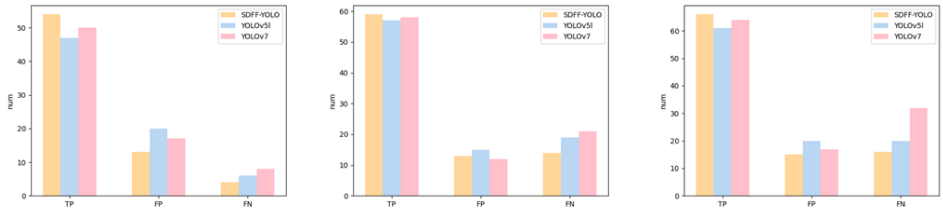

Figure 11. Distribution of indicators (TP, FP, and FN) for three methods (YOLOv5l, YOLOv7, and SMFF-YOLO) in three scenarios: (**a**) Scenario 1; (**b**) Scenario 2; (**c**) Scenario 3. The yellow bars represent the results of SMFF-YOLO, the blue bars represent the results of YOLOv5l, and the red bars represent the results of YOLOv7.

Analysis of Figure 11 reveals notable distinctions concerning true positives between the proposed method in this paper and other models across different scenes. These findings indicate that the proposed method achieves higher accuracy compared to other models. Additionally, our method exhibits clear advantages in the comparison of false positives and false negatives. Specifically, in Scenario 1 and Scenario 3, the SMFF-YOLO method demonstrates only half the number of false negatives compared to the baseline model (YOLOv7). These observations further affirm the superiority of our proposed SMFF-YOLO in object detection tasks involving UAV-captured images.

5. Conclusions and Future Work

In this paper, we proposed SMFF-YOLO, a scale-adaptive YOLO framework, to address the precise detection of multi-scale and tiny objects in UAV-captured images. Our framework introduced several key improvements. Firstly, we improved the detection performance for tiny objects by designing new prediction head modules and adding an additional head dedicated to tiny object detection. By merging Swin Transformer with CNN, we effectively leveraged global context and local features, resulting in enhanced accuracy. Secondly, we introduced the BFFP module in the neck part, which employs a bidirectional fusion strategy to enhance low-level information in the feature map. Finally, we designed the AASPP module to address the challenge of complex backgrounds in UAV-captured images. This module utilizes hybrid attention and cascaded atrous convolutions to extract multi-scale feature information, adapt to different target scales, and enhance the detection accuracy of multi-scale objects. Extensive experiments conducted on the VisDrone and UAVDT datasets demonstrated that SMFF-YOLO achieves higher accuracy compared to other methods. Furthermore, it exhibits robustness in challenging scenarios characterized by complex backgrounds, tiny objects, and occluded targets. In summary, SMFF-YOLO has made significant advancements in accurately detecting multi-scale objects and tiny objects in UAV-captured images.

Our method exhibited outstanding results in UAV scenarios. However, in challenging conditions marked by low lighting or fog, our method encountered difficulties in object detection. Therefore, we are considering the adoption of image enhancement along with lightweight strategies to improve its overall capability and applicability. Furthermore, our future research involves expanding our technique to a wider range of modalities, including infrared, SAR, and hyperspectral imagery.

Author Contributions: Conceptualization, H.Z. and X.Z.; methodology, Y.W. and H.Z.; software, H.Z. and X.Z.; validation, H.Z., Y.W. and M.Y.; formal analysis, Y.W., M.Y. and H.Z.; data curation, Y.W. and H.Z.; writing—original draft preparation, Y.W. and H.Z.; writing—review and editing, Y.W. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Bingtuan Science and Technology Program (grant no. 2019BC008) and in part by the National Natural Science Foundation of China under grant no. U1903214.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank the authors of the comparative methods, including YOLOv5, YOLOv7, and Swin Transform. Our deepest gratitude goes to the reviewers and editors for their careful work and thoughtful suggestions that have helped improve this paper substantially.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gu, J.; Su, T.; Wang, Q.; Du, X.; Guizani, M. Multiple moving targets surveillance based on a cooperative network for multi-UAV. *IEEE Commun. Mag.* **2018**, *56*, 82–89. [CrossRef]
2. Hird, J.N.; Montaghi, A.; McDermid, G.J.; Kariyeva, J.; Moorman, B.J.; Nielsen, S.E.; McIntosh, A.C. Use of unmanned aerial vehicles for monitoring recovery of forest vegetation on petroleum well sites. *Remote Sens.* **2017**, *9*, 413. [CrossRef]
3. 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, QC, Canada, 7–12 December 2015; Volume 28.
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
5. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
6. 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; pp. 2980–2988.
7. 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; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
8. Li, F.; Zeng, A.; Liu, S.; Zhang, H.; Li, H.; Zhang, L.; Ni, L.M. Lite DETR: An interleaved multi-scale encoder for efficient detr. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18558–18567.
9. 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 Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
10. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
11. Zhu, L.; Xiong, J.; Xiong, F.; Hu, H.; Jiang, Z. YOLO-Drone: Airborne real-time detection of dense small objects from high-altitude perspective. *arXiv* 2023, arXiv:2304.06925.
12. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. BiFormer: Vision Transformer with Bi-Level Routing Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
13. Li, C.; Zhou, A.; Yao, A. Omni-dimensional dynamic convolution. *arXiv* **2022**, arXiv:2209.07947.
14. Qi, G.; Zhang, Y.; Wang, K.; Mazur, N.; Liu, Y.; Malaviya, D. Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion. *Remote Sens.* **2022**, *14*, 420. [CrossRef]
15. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
16. Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. How far are we from solving pedestrian detection? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1259–1267.
17. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. Reppoints: Point set representation for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9657–9666.
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.

19. 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/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25.
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
23. 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; pp. 2961–2969.
24. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
25. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 821–830.
26. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
27. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
28. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
29. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
30. 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. [CrossRef] [PubMed]
31. Wu, H.; Hua, Y.; Zou, H.; Ke, G. A lightweight network for vehicle detection based on embedded system. *J. Supercomput.* **2022**, *78*, 18209–18224. [CrossRef]
32. Chen, Y.; Li, J.; Niu, Y.; He, J. Small object detection networks based on classification-oriented super-resolution GAN for UAV aerial imagery. In Proceedings of the 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 3–5 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4610–4615.
33. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered object detection in aerial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8311–8320.
34. Zhang, R.; Newsam, S.; Shao, Z.; Huang, X.; Wang, J.; Li, D. Multi-scale adversarial network for vehicle detection in UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 283–295. [CrossRef]
35. Guo, Y.; Chen, S.; Zhan, R.; Wang, W.; Zhang, J. LMSD-YOLO: A Lightweight YOLO Algorithm for Multi-Scale SAR Ship Detection. *Remote Sens.* **2022**, *14*, 4801. [CrossRef]
36. Sun, Z.; Leng, X.; Lei, Y.; Xiong, B.; Ji, K.; Kuang, G. BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images. *Remote Sens.* **2021**, *13*, 4209. [CrossRef]
37. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
38. Zhao, Q.; Liu, B.; Lyu, S.; Wang, C.; Zhang, H. TPH-YOLOv5++: Boosting Object Detection on Drone-Captured Scenarios with Cross-Layer Asymmetric Transformer. *Remote Sens.* **2023**, *15*, 1687. [CrossRef]
39. Zhang, Z.; Lu, X.; Cao, G.; Yang, Y.; Jiao, L.; Liu, F. ViT-YOLO: Transformer-based YOLO for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2799–2808.
40. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021; pp. 10781–10790.
41. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021; pp. 11534–11542.
42. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
43. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
44. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
45. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7036–7045.

46. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516.
47. Qiao, S.; Chen, L.C.; Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021; pp. 10213–10224.
48. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
49. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. *AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
50. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
51. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
52. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
54. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
55. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
56. Zhang, R.; Shao, Z.; Huang, X.; Wang, J.; Wang, Y.; Li, D. Adaptive dense pyramid network for object detection in UAV imagery. *Neurocomputing* **2022**, *489*, 377–389. [CrossRef]
57. Zhang, R.; Shao, Z.; Huang, X.; Wang, J.; Li, D. Object detection in UAV images via global density fused convolutional network. *Remote Sens.* **2020**, *12*, 3140. [CrossRef]
58. Zhang, J.; Huang, J.; Chen, X.; Zhang, D. How to fully exploit the abilities of aerial image detectors. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
59. 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.* **2020**, *30*, 1556–1569. [CrossRef]
60. Li, C.; Yang, T.; Zhu, S.; Chen, C.; Guan, S. Density map guided object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 190–191.
61. 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/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 3258–3267.
62. 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/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2789–2798.
63. Huang, Y.; Chen, J.; Huang, D. UFPMP-Det: Toward accurate and efficient object detection on drone imagery. *AAAI Conf. Artif. Intell.* **2022**, *36*, 1026–1033.