remote sensing

MDPI

Article

A Feature-Enhanced Anchor-Free Network for UAV Vehicle Detection

Jianxiu Yang [1,2], Xuemei Xie [1,*], Guangming Shi [1] and Wenzhe Yang [1]

[1] School of Artificial Intelligence, Xidian University, Xi'an 710071, China; jianxiuyang@stu.xidian.edu.cn (J.Y.); gmshi@xidian.edu.cn (G.S.); wzyang@stu.xidian.edu.cn (W.Y.)
[2] School of Physics and Electronics, Shanxi Datong University, Datong 037009, China
* Correspondence: xmxie@mail.xidian.edu.cn

Received: 29 July 2020; Accepted: 18 August 2020; Published: 24 August 2020

Abstract: Vehicle detection based on unmanned aerial vehicle (UAV) images is a challenging task. One reason is that the objects are small size, low-resolution, and large scale variations, resulting in weak feature representation. Another reason is the imbalance between positive and negative examples. In this paper, we propose a novel architecture for UAV vehicle detection to solve above problems. In detail, we use anchor-free mechanism to eliminate predefined anchors, which can reduce complicated computation and relieve the imbalance between positive and negative samples. Meanwhile, to enhance the features for vehicles, we design a multi-scale semantic enhancement block (MSEB) and an effective 49-layer backbone which is based on the DetNet59. The proposed network offers appropriate receptive fields that match the small-sized vehicles, and involves precise localization information provided by the contexts with high resolution. The MSEB strengthens discriminative feature representation at various scales, without reducing the spatial resolution of prediction layers. Experiments show that the proposed method achieves the state-of-the-art performance. Particularly, the main part of vehicles, much smaller ones, the accuracy is about 2% higher than other existing methods.

Keywords: feature-enhanced; anchor-free network; multi-scale; unmanned aerial vehicle; object detection

1. Introduction

Vehicle detection in unmanned aerial vehicle (UAV) images has received significant attention due to its extensive applications in both military and civilian fields, such as disaster management [1], transportation surveillance [2–4], and smart parking [5]. However, UAV vehicle detection is a challenging task because of small-sized objects, low-resolution objects, large object scale variations (e.g., large truck is about 460×300 pixels, while small bicycle is only about 20×20 pixels on 1920×1080 image. It can be seen that the scale between the large truck and the small bicycle is quite different, causing the problem of large scale variations still exists on UAV aerial images.), and the imbalance between positive and negative examples. Therefore, how to accurately and quickly detect vehicles in UAV images has theoretical significance and practical application value.

Traditional vehicle detection in UAV images is mainly based on hand-crafted features followed by a classifier or cascade of classifiers within a sliding window [6–10]. The hand-crafted features are low-level with weak semantics, which cannot represent vehicles effectively. Recently, thanks to the powerful representation capability of deep convolutional neural networks (CNN), object detection [11–15] has been made significant breakthroughs in the classical types of images (ground-level images), which also inspires vehicle detection in UAV images.

Many CNN-based vehicle detectors can be grouped into two main streams: the two-stage vehicle detectors and single-stage vehicle detectors. The two-stage vehicle detectors [16–19] are based on deep learning detection framework, such as Fast R-CNN [20], Faster R-CNN [12], which have considered enhancing the feature representation ability by introducing useful contextual information around vehicles. These methods can guarantee high accuracy but are not suitable for real-time applications. The single-stage vehicle detectors [21–24], such as YOLOv3-based [25] and RefineDet-based [26], can realize real-time detection and use top-down architecture [27,28] to enhance the feature representation for vehicles. These methods leverage anchor boxes with predesigned scales and aspect ratios to predict vehicles of different sizes. However, even with careful design, because the scales and aspect ratios of anchors are fixed for these anchor-based vehicle detectors, it is difficult to deal with object candidates with large scale variations, particularly for small-sized vehicles. Moreover, during training, all anchor boxes must be computed the intersection-over-union (IoU) [29] scores with all ground-truth boxes, resulting in an imbalance between positive and negative samples, a large amount of computation and memory footprint.

In this paper, we adopt the anchor-free mechanism [30] to eliminate predefined anchor boxes, avoiding excessive complex related calculations and handling large scale variations. The proposed feature-enhanced anchor-free (FEAF) network for UAV vehicle detection, uses a fully convolutional neural network (FCN) to directly output the pixel-wise classification scores and the object bounding boxes. The FEAF network is the same as FCN [31], which views each location of the vehicle bounding boxes as positive samples. Compared to anchor-based vehicle detectors that only consider the anchor boxes with a highly enough IoU with ground-truth boxes as positive samples, the FEAF detector has more positive samples to train, especially for small-sized vehicles. Therefore, the proposed FEAF network can relieve the imbalance between positive and negative samples to achieve better performance.

Furthermore, UAV vehicles are small size and low-resolution, resulting in weak feature representation. Therefore, we design an effective backbone and a multi-scale block to enhance the features for vehicles. On one hand, most object detectors [32–36] use ResNet [37,38] or ResNext [39] as the backbone network and use large down-sampling factors (e.g., 32, 64, and 128) in feature maps to build a deeper pyramid network. DetNet [40–43] indicates that large down-sampling factors can bring large valid receptive field, which is good for image classification but will impair the object location ability. Thus, small down-sampling factors can keep precise location information, which is beneficial to object location. On the other hand, the proposed FEAF network uses each location to directly regress the target bounding box. Obviously, precise location information is even more necessary than the anchor-based detectors. Therefore, we use an effective 49-layer backbone based on the DetNet59 [40,41], which maintains precise localization information with high spatial resolution in deep convolutional layers. Moreover, the proposed backbone includes fewer layers against deeper layers to offer matched receptive fields for small-sized vehicles.

Apart from this, we propose a multi-scale semantic enhancement block (MSEB) to strengthen discriminative feature representation. Generally, the deeper layers have stronger semantic information, but localization will suffer from the absence of the fine location information. Meanwhile, the deeper layers create large receptive field accompanying background interference for small and low-resolution vehicles, which will limit the vehicle location ability. The proposed MSEB will widen the network width instead of increasing the depth, which can effectively enhance the semantics for vehicles at various scales, without changing the spatial resolution of prediction layers.

As a summary, we have made the following main contributions:

(1) We propose a feature-enhanced anchor-free network (FEAF) for UAV vehicle detection, reducing excessive complex calculations related to anchor boxes and relieving the imbalance between positive and negative samples.
(2) We adopt an effective 49-layer backbone that can offer appropriate receptive fields and keep precise localization information to match exactly small-sized vehicles. Besides, a multi-scale semantic enhancement block (MSEB) is proposed to strengthen discriminative feature representation for vehicles at various scales, without changing the spatial resolution of prediction layers.
(3) Our method achieves the state-of-the-art performance on the two datasets, which are the UAVDT dataset [44] and the XDUAV dataset [45]. On the first dataset, 81.4% AP is achieved, which is 2.4%, 1.7%, 0.8%, and 2.2% higher than FPN [28], Mask R-CNN [15], FCOS [30], and RetinaNet [32] respectively. On the second one, 73.5% AP is achieved, which is 1.6%, 1.7%, 1.9%, and 2.5% higher than FPN, Mask R-CNN, FCOS, and RetinaNet respectively. Particularly, the main part of UAV datasets is much smaller vehicles, and its accuracy AP_S is about 2% higher than other existing methods. While, the proposed detector can run at 22 frames per second on a single NVIDIA TITAN Xp GPU.

The organization of the rest part of this paper is as follows. Section 2 introduces some related works of this paper. Section 3 describes the technical design and theoretical analysis of the proposed method. Section 4 presents experimental results of the proposed method and gives detailed analysis. Section 5 draws the conclusion.

2. Related Work

2.1. Anchor-Based UAV Vehicle Detection

With the impressive progress of deep learning in image processing, UAV vehicle detectors based convolutional neural networks (CNNs) have been proposed in recent years. Many UAV vehicle detection methods [19,46] based two-stage detectors, which can achieve improved performance by enhanced feature representation for one category vehicle. Sommer et al. [47] exploit Faster R-CNN [12] to extend the detection task for multiple vehicle categories. Zhang et al. [48] adopt Cascade R-CNN [49] to realize dense and small vehicles detection in UAV vision. However, these UAV vehicle detection methods cannot satisfy real-time requirement. Subsequently, the single-stage detectors can achieve real-time object detection. Radovic et al. [21] and Tang et al. [22] use YOLO [50] and YOLOv2 [51] to complete fast vehicle detection and tracking in UAV images, respectively. ShuffleDet [52] applies inception modules and deformable modules to consider the size and geometric shape of the vehicles to finish real-time vehicle detection. However, these vehicle detectors rely on pre-defined anchor boxes that bring an imbalance between positive and negative samples, a large amount of computation and memory footprint.

2.2. Anchor-Free UAV Vehicle Detection

To solve the problems caused by setting anchor boxes, anchor-free detectors [53–61] are proposed to eliminate pre-designed scales and aspect ratios of anchors and directly output the bounding boxes from an image. CornerNet [62] directly detects a pair of corners of an object bounding box and groups them via associative embedding [63] technique. CornerNet achieves superior performance but comes at high post-processing cost. FCOS [30] and FoveaBox [64] consider locations located the ground-truth box as positives to predict four distances. For UAV vehicles detection, RRNet [65] first uses the anchor-free detector to generate the coarse boxes, and then applies a re-regression module to produce accurate bounding boxes. Cai et al. [66] propose an anchor-free Guided Attention Network (GANet) to deal with object detection and counting tasks, but it only achieves one class UAV vehicle

detection. In this paper, we use anchor-free mechanism to achieve efficiency and effective multi-class vehicle detection in UAV images.

2.3. Feature-Enhanced UAV Vehicle Detection

Vehicles in UAV images are small size and low-resolution, resulting in weak feature representation of vehicle targets. Feature enhancement is very necessary for vehicle detection. Intuitively, contextual information [67–71] is helpful for small and low-resolution objects. Many vehicle detectors in UAV images [72,73] use the FPN-based network to offer contextual information, which realize high accuracy vehicle detection and counting tasks. SlimYOLOv3 [74] uses fewer trainable parameters and floating point operations (FLOPs) by pruning feature channels to achieve real-time vehicle detection. Ammar et al. [75] present a performance evaluation between Faster R-CNN [12] and YOLOv3 [25] for car detection in UAV images. Liang et al. [76] use feature fusion and scaling-based SSD for small vehicles detection in UAV images. In this paper, we use an effective backbone that maintains high spatial resolution in deeper convolutional layers. Thus, the contextual information is generated through top-down and lateral connection in the network, which can introduce precise localization information to enhance the features for small-sized vehicles. Moreover, a multi-scale semantic enhancement block (MSEB) is proposed to strengthen discriminative feature representation for vehicles at various scales.

3. Proposed Method

In this section, we elaborate the proposed feature-enhanced anchor-free network (FEAF) for UAV vehicle detection. The overall architecture is shown in Figure 1.

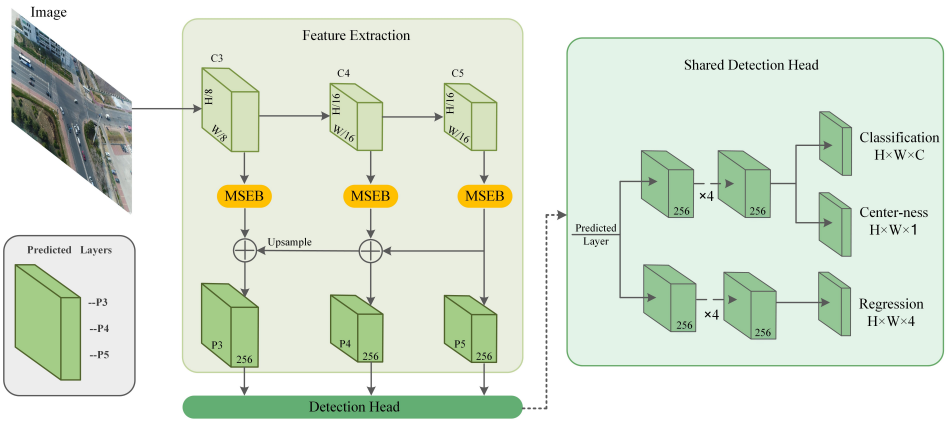


Figure 1. The over architecture of the FEAF network, including feature extraction and shared detection head. The MSEB is a multi-scale semantic enhancement block to strengthen semantic information.

3.1. Architecture

We introduce the proposed architecture for UAV vehicle detection in detail, including feature extraction and detection head.

Feature Extraction. The proposed FEAF network adopts the anchor-free mechanism to achieve regression. If the location (x, y) falls into any ground-truth boxes, it can be considered as a positive sample to directly output the pixel-wise classification scores and the object bounding boxes. Therefore, it is very important to maintain accurate location information for the FEAF network. The DetNet59 [40] employs small down-sampling factors (i.e., 4, 8, and 16) to maintain fine location information while employing dilated convolution [77] to increase receptive field. Inspired by this, we adopt an effective 49-layer backbone based on DetNet59 to extract features as shown in Figure 1. H and W in Figure 1 are height and width of feature maps (e.g., C_3, C_4, and C_5) receptively, '/d' (d = 8, 16, 16) is the down-sampling factor of feature levels to the input image. We use the

convolutional stages from C_1 to C_5 as the backbone called DetNet-49 and remove the deeper convolutional layer stage6 from the DetNet59. Because the deeper convolutional layers introduce large receptive field for small-sized vehicles, it will bring too much background inference to compromise the detection performance. Thus, the proposed backbone DetNet-49 can maintain precise localization information in deeper convolutional layers while providing matched receptive fields for small-sized vehicles.

The top-down architecture with lateral connection is used to build a feature pyramid structure from P_3 to P_5, and offer contextual information to enhance the features of vehicles. The contextual information involves semantic information and more precise localization information with high spatial resolution. The predicted layers P_3, P_4, and P_5 are produced by the feature maps C_3, C_4, and C_5 followed by a multi-scale semantic enhancement block (MSEB) with lateral connection. The MSEB will be explained in detail in Section 3.2. Meanwhile, the channel dimensions of predicted layers are fixed at 256 as shown in Figure 1, further reducing the computational cost.

Detection Head. Many literatures [60,78] have extensively explored that the detection head plays an important role in high performance. Same as FCOS [30] and RetinaNet [32], we append four 3×3 convolutional layers on the detection head respectively for classification and regression branches to improve detection accuracy. Meanwhile, we employ a center-ness branch [30], in parallel with the classification branch to suppress these low-quality predicted bounding boxes without introducing any hyper-parameters, which will be introduced in detail in Section 3.3. The parameters of the head are shared across all predicted levels. For UAV vehicles detection, we predict a C dimensional (C is the number of categories) vector $\boldsymbol{p}$ of classification labels and a 4 dimensional vector $\boldsymbol{t}$ of bounding box coordinates in the final layers. In shared detection head, all parameters are shared across predicted levels, and four 3×3 convolutional layers are added respectively for classification and regression branches. The center-ness branch decreases low-quality predicted bounding boxes.

3.2. Multi-Scale Semantic Enhancement Block

Although the network composed of only the DetNet-49 is conducive to the detection for small-sized vehicles, it is not good for classification and regression of large-sized vehicles. Intuitively, the deeper network creates large receptive fields and stronger semantic information that are good for large-sized vehicles classification. However, localization will suffer from the absence of the fine location information, and large receptive fields will accompany background interference for small-sized and low-resolution vehicles. In order to maintain precise location information and offer matched receptive fields for all-sized vehicles, we will widen the network width instead of increasing the depth, which can effectively strengthen the semantics of vehicles at various scales. Based on the DetNet-49 backbone, we design a multi-scale semantic enhancement block (MSEB) as shown in Figure 2 to widen the network, without changing the spatial resolution of prediction layers.

The MSEB contains three branches that are stacked by convolution kernels of different size. Specifically, the first branch of the MSEB decomposes a 3×3 convolution kernel into a 1×3 and a 3×1 kernels, the last two branches use a 3×3 convolutional kernel followed by a 1×1 kernel and a 1×1 convolutional kernel respectively. We fix the number of channels in each layer of the MSEB to 256. Finally, the three branches are summed in an element-wise manner. Furthermore, we employ batch-normalization (BN) [79] after each convolutional layer in the MSEB. In this way, the proposed network can not only strengthen semantic information for vehicles, but also maintain precise location information and create various receptive fields with different scales from one feature map.

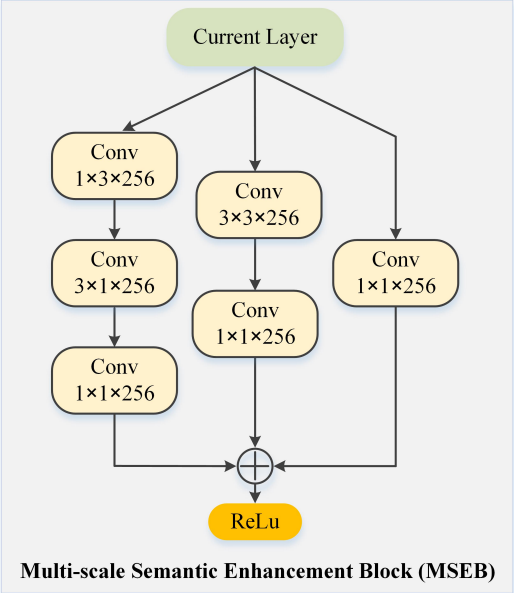


Figure 2. The proposed multi-scale semantic enhancement block (MSEB).

3.3. Anchor-Free Mechanism

In this paper, we adopt the anchor-free mechanism [30] to regress bounding boxes of UAV vehicles. Given a ground-truth box denoted as (x_t, y_t, x_b, y_b) , the points (x_t, y_t) and (x_b, y_b) refer to the left-top and right-bottom corners of the target bounding box respectively. Each location (x, y) on the predicted layer P_i can be mapped back onto the input image $(\lfloor s/2 \rfloor + xs, \lfloor s/2 \rfloor + ys)$, which is near the center of the receptive field of the location (x, y). s is the total stride and is 8, 16, and 16 in P_3, P_4, and P_5 respectively. If the location (x, y) falls into any ground-truth boxes, it can be considered as a positive sample; otherwise it is a negative one. If a location falls into multiple ground-truth boxes, we adopt the multi-level prediction to solve this problem. Besides the label for classification, the detection head directly regresses a 4 dimensional vector for each location on all predicted layers that represents the distances between the current location and the four bounds of the ground-truth box. For example, if location (x, y) falls into the given a ground-truth box (x_t, y_t, x_b, y_b), the regression targets for the location can be calculated as a 4D vector $\boldsymbol{t}^*_{x,y}$,

$$\boldsymbol{t}^*_{x,y} = \{dx_t = x - x_t, dy_t = y - y_t, dx_b = x_b - x, dy_b = y_b - y\}. \tag{1}$$

If a location satisfies $\max(dx_t, dy_t, dx_b, dy_b) > m_i$ or $\max(dx_t, dy_t, dx_b, dy_b) < m_{i-1}$, we take it as a negative sample and is not required to regress it, where m_i is the maximum distance that the predicted layer P_i needs to regress. In this paper, m_3, m_4, and m_5 are set as 64, 256, and ∞ respectively. In other words, the size range is [0, 64] for P_3, [64, 256] for P_4, and [256, ∞] for P_5. Therefore, compared to the anchor-based detectors, the anchor-free mechanism can make full use of more positive samples to train. Since the anchor-based detectors only consider the anchor boxes with larger IoU as positive samples.

Center-ness. Since the positive samples far away from the target center will regress low-quality bounding boxes, a constraint called the center-ness strategy is used to remove those boxes. For a regression box (dx_t, dy_t, dx_b, dy_b) of the location (x, y), the center-ness is defined as,

$$(cn)^* = \sqrt{\frac{\min(dx_t, dx_b)}{\max(dx_t, dx_b)} \times \frac{\min(dy_t, dy_b)}{\max(dy_t, dy_b)}}. \tag{2}$$

The center-ness does not introduce other hyper-parameters and without fine-tuning in training process. During the testing, the final scores used for ranking the bounding boxes can be calculated by multiplying the predicted center-ness with the corresponding classification confidence. Thus the weights of the bounding boxes far away from the center point are smaller. Therefore, non-maximum suppression (NMS) can filter out those low-quality boxes and improve UAV vehicles detection performance.

Loss Function. During the training phase, we minimize an objective function following the multi-task loss,

$$L(\boldsymbol{p}_{x,y}, \boldsymbol{t}_{x,y}, \boldsymbol{cn}) = \frac{1}{N_{pos}} \sum_{(x,y)} L_{cls}(\boldsymbol{p}_{x,y}, c^*_{x,y}) + \frac{1}{N_{pos}} \sum_{(x,y)} c^*_{x,y} L_{cn}(\boldsymbol{cn}, \boldsymbol{cn}^*) + \frac{1}{N_{pos}} \sum_{(x,y)} c^*_{x,y} L_{reg}(\boldsymbol{t}_{x,y}, \boldsymbol{t}^*_{x,y}). \quad (3)$$

where the classification loss L_{cls} uses focal loss [32], the center-ness loss L_{cn} is binary cross entropy loss [12], and the regression loss L_{reg} adopts GIoU loss [80]. N_{pos} is the number of positive samples. The label $c^*_{x,y}$ is 1 if the location (x, y) is positive and 0 otherwise. The summation is calculated over all location on the pyramid level $P_i (i = 3, 4, 5)$.

4. Experimental Results and Analysis

In this section, we analyze the proposed network for vehicle detection. In the ablation study, backbone network, multi-scale semantic enhancement block (MSEB), and anchor-free mechanism are analyzed in detail.

4.1. Dataset Preparation and Training Implementation Details

XDUAV Dataset. The XDUAV dataset [45] was captured by the DJI Phantom 2 quadcopter flying a part of urban and suburban areas of Xi'an, China. The dataset consists of 11 vehicle videos from various traffic environment, such as congested and non-congested conditions, and intersection scenarios in different weather and lighting conditions. All videos are collected with the drone-view at approximate 100 meters' height. The resolution is 1920×1080 and we captured one target image per 30 frames, and the whole dataset contains 4344 images with 3475 images for training and 869 images for testing. The dataset has a large amount of small vehicles with truncated, occluded, and multi-angle. Figure 3a shows some image examples in different scenarios and weather conditions. Training and testing samples are annotated 6 categories of vehicles (i.e., car, bus, truck, tanker, motor and bicycle). Each category object number is shown in Table 1.

UAVDT Dataset. UAV Detection and Tracking (UAVDT) benchmark [44] was captured by DJI Inspire 2 flying different altitude in urban areas, under various weather and lighting conditions in different scenarios such as arterial streets, highways, crossing and T-junctions, etc. The UAVDT benchmark consists of about 80,000 representative frames for three fundamental tasks, i.e., object detection, single object tracking and multiple object tracking. In this paper, we only implement the fundamental task of object detection. The dataset for object detection contains 39,850 images with 23,258 images for training and 16,592 images for testing, with the resolution of 1080×540 pixels. The dataset contains 3 annotated categories including car, truck and bus. The vehicle number of each category and some images examples are shown in Table 1 and Figure 3b, respectively.

Metric. In this paper, we adopt MS COCO metric to evaluate the results of the proposed detector on two datasets. MS COCO metric can judge detection performance under a rigorous manner, including AP, AP_{50}, AP_{75}, AP_S, AP_M, AP_L metric. The mean average precision (AP) is calculated by averaging over all 10 Intersection over Union (IoU) thresholds in the range [0.5, 0.95] with an interval 0.05 of all categories. The AP value is the primary metric for ranking. AP_{50} and AP_{75} are calculated by the average of all categories at a single IoU value 0.5 and 0.75 respectively. Apart from that, the AP_S, AP_M, AP_L values are computed separately for small-sized (area $< 32^2$), medium-sized ($32^2 <$ area $< 96^2$) and large-sized (area $> 96^2$) objects in order to measure the detection performance on targets of different sizes.

(a) Examples on the XDUAV dataset

(b) Examples on the UAVDT dataset

Figure 3. Samples in various scenes from different vehicle datasets. (**a**) examples on the XDUAV dataset; (**b**) examples on the UAVDT dataset.

Table 1. The number of each category vehicle in the XDUAV dataset and the UAVDT dataset.

XDUAV Dataset	Category	Car	Bus	Truck	Motor	Bicycle	Tanker
	Number	33,841	2690	2848	6656	2024	173
UAVDT Dataset	Category	Car	Truck	Bus	-	-	-
	Number	755,197	25,086	17,450	-	-	-

Training Implementation Details. The proposed UAV vehicle detector FEAF is end-to-end trained on 4 NVIDIA TITAN Xp GPUs with a total of 16 images per minibatch (4 images per GPU). Our network is optimized by stochastic gradient descent (SGD) with a weight decay of 0.0001 and momentum of 0.9. Unless otherwise specified, all models are trained for 100 k iterations with an initial learning rate being 0.01, which is reduced by a factor of 10 at iteration 60 k and 80 k respectively. We initialize our backbone network using the pretrained weights on ImageNet [81]. We resize the shorter side of the input images to 800 and the longer side less or equal to 1333 to avoid too much memory cost.

4.2. Ablation Study

In this section, we take the XDUAV dataset as an example to conduct an ablative analysis for the proposed vehicle detector.

4.2.1. Backbone Network Analysis

To maintain precise spatial location information and demonstrate the effectiveness of the backbone DetNet-49 for small-sized vehicles detection, we make comparative experiments in different backbone networks, including ResNet-50, ResNet-101, and DetNet59 as shown in Table 2. All experiments are used anchor-free regression mechanism [30] to detect UAV vehicles.

Table 2. Performance comparisons between the proposed backbone DetNet-49 and other backbones in terms of MS COCO metric. The bold numbers represent the best results.

Methods	Backbone	Additional Layers	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS [30]	ResNet-50	w	71.2	95.4	82.3	33.2	68.1	80.9
	ResNet-50	w/o	70.2	95.4	80.9	34.1	65.3	79.3
	ResNet-101	w	71.6	96.1	82.6	34.1	67.1	81.3
	ResNet-101	w/o	71.3	96.1	82.4	34.3	68.3	80.2
Ours *	DetNet-59	-	71.6	96.0	83.1	35.1	67.9	81.0
	DetNet-49	-	**71.9**	95.7	**83.5**	**35.4**	**69.4**	80.1

* We use DetNet-59 and DetNet-49 as the backbone respectively to achieve anchor-free regression location.

Firstly, to maintain precise spatial location information, we adopt the DetNet59 [40] as the backbone that employs small downsampling factors (i.e., 4, 8, and 16). The FCOS [30] uses ResNet-50 and ResNet-101 as the backbone respectively, and increases two additional predicted layers using large downsampling factors 64 and 128. Line 2 and line 6 from Table 2 illustrate that DetNet59 based on anchor-free regression mechanism is 0.4% and 1.9% higher in AP and AP_S respectively than ResNet-50. Although the AP value is the same when the backbone is ResNet-101 and DetNet59 as shown in line 4 and line 6, the AP_S and AP_M value of DetNet59 as the backbone are higher than that of ResNet-101. For fair comparison, we remove two additional layers with the large factors 64 and 128 in FCOS [30]. Experimental results from Line 3 and Line 5 without additional layers show that the accuracy AP_S for small-sized vehicles is higher than that with additional layers, but at the same time it is lower than the accuracy of DetNet59 with the small factors (i.e., 8 and 16). These results evidence the importance of spatial location information for small-sized vehicles.

Secondly, to prove the effectiveness of the proposed DetNet-49, we analyze the role of a deeper convolutional layer stage6 from the DetNet59. As shown in line 6 and line 7 from Table 2, the comparison of experimental results proves the deeper convolutional layer stage6 is helpless for small-sized vehicle detection. Since the deeper convolutional layers introduce large receptive field for small-sized vehicles, it will bring too much background inference to compromise the detection performance. The experimental results show that the proposed backbone DetNet-49 not only reduces the parameters and the computational cost of the model to train more stable, but also offers precise location information and matched receptive fields to improve the detection performance for small-sized vehicles.

4.2.2. Multi-Scale Semantic Enhancement Block (MSEB) Analysis

As shown in Table 2, the AP_L value of DetNet-49 as the backbone is less than that of other backbones, which shows that the absence of deeper convolutional layers leads to insufficient semantics for large-sized vehicles. Therefore, we use the proposed MSEB to enhance the semantics for the large-sized vehicles while maintaining spatial location information for the small-sized vehicles.

In order to prove the effectiveness of the designed MSEB for vehicles, we adopt different blocks as shown in Figure 4 to predict the targets and analyze their impact on the detection performance. The original feature pyramid network (FPN) only uses a 1×1 convolutional layer that is named M_1 block to perform lateral connection as shown in Figure 4a. For the proposed FEAF, the M_1 block cannot enhance semantics for UAV vehicles. We adopt two 3×3 convolutional layers (M_2 block as shown in Figure 4b) instead of M_1 block to enhance vehicles semantics. The experimental results show that the AP and AP_L of the M_2 block are 0.5% and 1.1% higher respectively than the M_1 block as shown in Table 3. Although the M_2 block has achieved the semantic enhancement for large-sized vehicles, the AP_S value of small-sized vehicles has decreased, which may be caused by the absence of the precise location due to the deep convolutional operation.

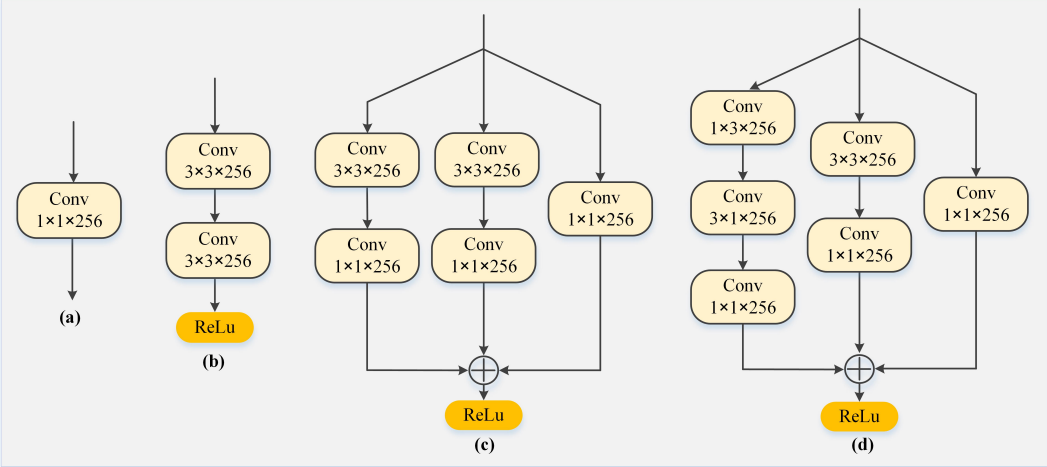


Figure 4. Different multi-scale semantic enhancement blocks. From left to right are (**a**) M_1 block, (**b**) M_2 block, (**c**) M_3 block, and (**d**) proposed MSEB block.

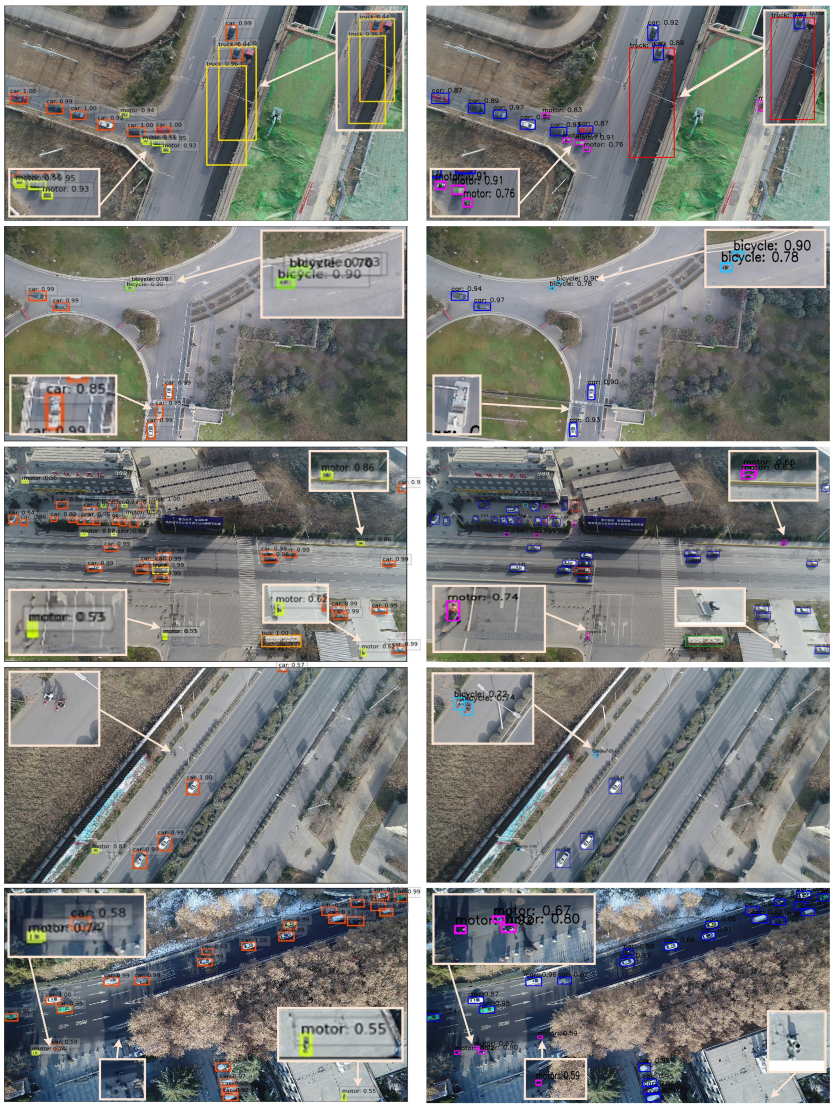
Table 3. Performance comparisons between the Proposed MSEB block and other blocks in terms of MS COCO metric. The bold numbers represent the best results.

Different Blocks	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
M_1 block	DetNet-49	71.9	95.7	83.5	35.4	69.4	80.1
M_2 block	DetNet-49	72.4	96.0	83.7	34.3	69.5	81.2
M_3 block	DetNet-49	72.6	95.7	83.7	34.8	69.1	81.3
Proposed MSEB block	DetNet-49	**73.5**	**97.5**	**84.2**	**35.7**	**71.9**	**81.4**

Inspired by the GoogleNet [82], we will widen the network width instead of increasing the depth as shown in Figure 4c,d, which can effectively improve the semantics of vehicles at various scales. Compared with the M_2 block, the M_3 block widens the network width to extract more rich semantics. However, the AP_S and AP_M of the M_3 block still become poor compared with the original M_1 block. To enhance the semantics and maintain location information for the small-sized vehicles, we decompose a 3×3 convolution kernel into 1×3 and 3×1 kernels as shown in Figure 4d. All AP values have been increased as shown in line 5 from Table 3, and the amount of parameters has also reduced. The proposed multi-scale semantic enhancement block (MSEB) can not only strengthen semantic information for vehicles, but also retain fine location information and create various receptive fields with different scales.

4.2.3. Anchor-Free vs. Anchor-Based Vehicle Detectors

The qualitative comparisons between the anchor-based and the proposed anchor-free methods can be seen in Figure 5. Compared with the anchor-based method [24] (Figure 5a), our anchor-free method (Figure 5b) obviously reduces false positives rate, missed rate, and redundant bounding boxes, etc. We analyze the two main reasons as following: firstly, because the scales and aspect ratios of anchors are fixed for the anchor-based vehicle detectors, even with careful design, it is difficult to deal with object candidates with large scale variations, particularly for small-sized vehicles. Secondly, anchor-based vehicle detectors that only consider the anchor boxes with a highly enough IoU with ground-truth boxes as positive samples, but the anchor-free methods consider the locations falling into any ground-truth boxes as positive samples. Compared with the anchor-based vehicle method, the anchor-free vehicle detector has more positive samples for training to reduce the imbalance between positive and negative samples.



(a) Detection results from anchor-based method (b) Detection results from our anchor-free method

Figure 5. Qualitative results comparing the proposed anchor-free method with the anchor-based method. (**a**) detection results from anchor-based method and (**b**) detection results from our anchor-free method.

4.3. Overall Performance

Due to the designs of the anchor-free regression mechanism, the effective backbone DetNet-49, and multi-scale semantic enhancement block, the proposed FEAF network can achieve 73.5% accuracy on the XDUAV dataset. We compare single-stage detection methods (including anchor-based methods and anchor-free methods) and two-stage high-accuracy detection methods by COCO metric as shown in Table 4. Among single-stage methods, the proposed FEAF obtains the top AP, which is even better than two-stage methods. Our method is 1.6%, 1.7%, 1.9%, and 2.5% higher in AP than FPN [28], Mask R-CNN [15], FCOS [30], and RetinaNet [32] respectively. All methods are trained under the same conditions, so the experimental results are credible.

To further demonstrate the robustness of the proposed method, we also evaluate it on the UAVDT dataset. In the analysis of Backbone Network, Multi-scale Semantic Enhancement Block (MSEB), and Anchor-free mechanism, we conduct the same ablative experiments on the XDUAV dataset. The experimental results of the UAVDT dataset can also prove the effectiveness of the proposed detector. Table 5 shows an overall result of different detection methods on the UAVDT

dataset. Our method achieves 81.4% AP that is 2.4%, 1.7%, 0.8%, and 2.2% higher than FPN [28], Mask R-CNN [15], FCOS [30], and RetinaNet [32] respectively. Figure 6 shows detection results on the UAVDT dataset in different scenarios using our proposed method. It is noted that the UAVDT dataset only focuses on vehicles on the main road, and vehicles parked around the road are ignored.

Table 4. Detection Results (%) of Different Methods for the XDUAV Dataset. The bold numbers represent the best results.

Methods	Backbone	Input Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
two-stage								
Faster R-CNN+++ [37]	ResNet-101	~1333 × 800	69.9	95.0	79.7	34.2	66.7	79.3
Faster R-CNN w FPN [28]	ResNet-101	~1333 × 800	71.9	96.8	83.7	35.0	67.4	80.5
Mask R-CNN [15]	ResNet-101	~1333 × 800	71.8	96.3	82.8	35.2	67.2	81.3
single-stage								
RetinaNet [32]	ResNet-101	~1333 × 800	71.0	95.1	81.5	31.8	67.1	81.4
FCOS [30]	ResNet-50	~1333 × 800	71.2	95.4	82.3	33.2	68.1	80.9
FCOS [30]	ResNet-101	~1333 × 800	71.6	96.1	82.6	34.1	67.1	81.3
Our Method	DetNet-49	~1333 × 800	**73.5**	**97.5**	**84.2**	**35.7**	**71.9**	**81.4**

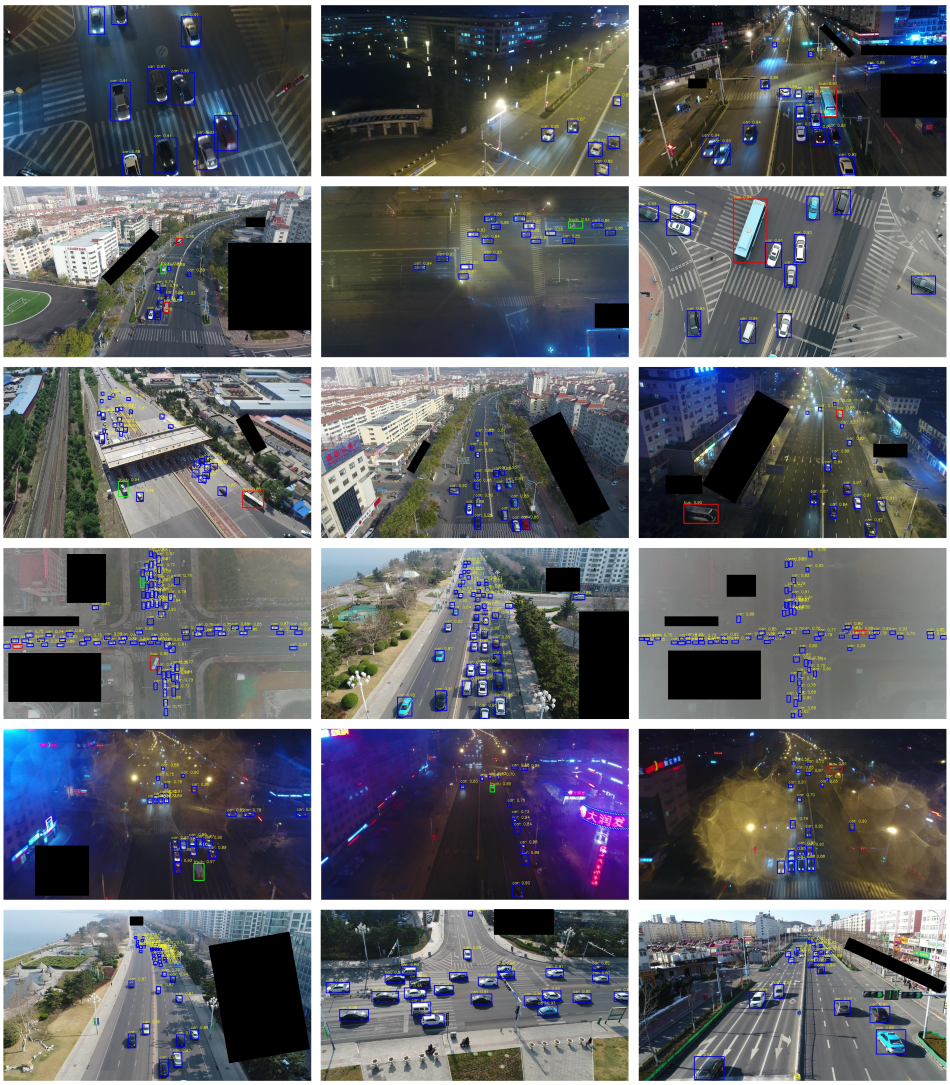


Figure 6. Qualitative detection results in different challenging scenarios. The black regions are ignored in the UAVDT dataset.

Table 5. Detection Results (%) of Different Methods for the UAVDT Dataset. The bold numbers represent the best results.

Methods	Backbone	Input Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
two-stage								
Faster R-CNN+++ [37]	ResNet-101	∼1333 × 800	74.1	98.1	88.7	67.4	82.3	89.8
Faster R-CNN w FPN [28]	ResNet-101	∼1333 × 800	79.0	97.9	93.3	73.1	85.5	90.7
Mask R-CNN [15]	ResNet-101	∼1333 × 800	79.7	98.3	94.2	74.1	86.0	90.7
single-stage								
RetinaNet [32]	ResNet-101	∼1333 × 800	79.2	97.6	90.4	71.0	86.8	93.7
FCOS [30]	ResNet-50	∼1333 × 800	79.4	98.7	93.2	73.1	85.9	93.5
FCOS [30]	ResNet-101	∼1333 × 800	80.6	98.7	93.2	73.3	87.1	93.7
Our Method	DetNet-49	∼1333 × 800	**81.4**	**98.7**	**94.5**	**75.3**	**87.4**	93.6

5. Conclusions

This paper proposes a novel feature-enhanced anchor-free network for vehicle detection in unmanned aerial vehicle (UAV) vision, which achieves accurate detection. Firstly, in order to avoid the problems caused by the setting anchors, we adopt the anchor-free mechanism to eliminate predefined anchor boxes, relieving the imbalance between positive and negative samples, and handling objects large scale variations. Secondly, to enhance the features for vehicles, we design a multi-scale semantic enhancement block (MSEB) and an effective backbone DetNet49. The backbone includes fewer layers against deeper layers to offer matched receptive fields for small-sized vehicles and precise localization information. The MSEB widens the network width, which can effectively strengthen discriminative feature representation of vehicles at various scales. The experimental results on two publicly available vehicle datasets demonstrate that the proposed method can achieve the state-of-the-art detection performance, which proves the effectiveness and robustness of the detector. In our future work, we intend to integrate logical reasoning relationships and knowledge priors into the vehicle detection network to further improve detection accuracy for more low-resolution vehicles.

Author Contributions: J.Y. and W.Y. contributed to the idea and the data collection of this study; J.Y. developed the algorithm, performed the experiments, analyzed the experimental results and worte this paper; X.X. and G.S. supervised the study and reviewed this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) under Grant 61836008 and Grant 61632019.

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

Abbreviations

UAV	unmanned aerial vehicle
CNN	convolutional neural networks
FCN	fully convolutional neural network
FEAF	feature-enhanced anchor-free network
MSEB	multi-scale semantic enhancement block
FPN	feature pyramid network
AP	mean average precision

References

1. Alotaibi, E.T.; Alqefari, S.S.; Koubaa, A. LSAR: Multi-UAV Collaboration for Search and Rescue Missions. *IEEE Access* **2019**, *7*, 55817–55832. [CrossRef]
2. Koubaa, A.; Qureshi, B. DroneTrack: Cloud-Based Real-Time Object Tracking Using Unmanned Aerial Vehicles Over the Internet. *IEEE Access* **2018**, *6*, 13810–13824. [CrossRef]

3. Leitloff, J.; Rosenbaum, D.; Kurz, F.; Meynberg, O.; Reinartz, P. An operational system for estimating road traffic information from aerial images. *Remote Sens.* **2014**, *6*, 11315–11341. [CrossRef]
4. Benjdira, B.; Khursheed, T.; Koubaa, A.; Ammar, A.; Ouni, K. Car Detection using Unmanned Aerial Vehicles: Comparison between Faster R-CNN and YOLOv3. In Proceedings of the 2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS), Muscat, Oman, 5–7 February 2019; pp. 1–6.
5. Li, X.; Chuah, M.C.; Bhattacharya, S. UAV assisted smart parking solution. In Proceedings of the 2017 International Conference on Unmanned Aircraft Systems (ICUAS), Miami, FL USA, 13–16 June 2017; pp. 1006–1013.
6. Liu, K.; Mattyus, G. Fast multiclass vehicle detection on aerial images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1938–1942.
7. Moranduzzo, T.; Melgani, F. Automatic car counting method for unmanned aerial vehicle images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 1635–1647. [CrossRef]
8. Tuermer, S.; Kurz, F.; Reinartz, P.; Stilla, U. Airborne vehicle detection in dense urban areas using HoG features and disparity maps. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 2327–2337. [CrossRef]
9. Moranduzzo, T.; Melgani, F. Detecting cars in UAV images with a catalog-based approach. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6356–6367. [CrossRef]
10. Teutsch, M.; Kruger, W. Robust and fast detection of moving vehicles in aerial videos using sliding windows. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 26–34.
11. 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]
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.
13. 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; pp. 21–37.
14. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *arXiv* **2016**, arXiv:1605.06409.
15. 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.
16. Xu, Y.; Yu, G.; Wang, Y.; Wu, X.; Ma, Y. Car detection from low-altitude UAV imagery with the faster R-CNN. *J. Adv. Transp.* **2017**, *2017*. [CrossRef]
17. Sommer, L.W.; Schuchert, T.; Beyerer, J. Fast deep vehicle detection in aerial images. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 311–319.
18. Koga, Y.; Miyazaki, H.; Shibasaki, R. A CNN-Based Method of Vehicle Detection from Aerial Images Using Hard Example Mining. *Remote Sens.* **2018**, *10*, 124. [CrossRef]
19. Tang, T.; Zhou, S.; Deng, Z.; Zou, H.; Lei, L. Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. *Sensors* **2017**, *17*, 336. [CrossRef] [PubMed]
20. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
21. Radovic, M.; Adarkwa, O.; Wang, Q. Object Recognition in Aerial Images Using Convolutional Neural Networks. *J. Imaging* **2017**, *3*, 21. [CrossRef]
22. Tang, T.; Deng, Z.; Zhou, S.; Lei, L.; Zou, H. Fast vehicle detection in UAV images. In Proceedings of the 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 19–21 May 2017; pp. 1–5.
23. Ringwald, T.; Sommer, L.; Schumann, A.; Beyerer, J.; Stiefelhagen, R. UAV-Net: A Fast Aerial Vehicle Detector for Mobile Platforms. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 544–552.
24. Yang, J.; Xie, X.; Yang, W. Effective Contexts for UAV Vehicle Detection. *IEEE Access* **2019**, *7*, 85042–85054. [CrossRef]
25. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.

26. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-Shot Refinement Neural Network for Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.
27. Shrivastava, A.; Sukthankar, R.; Malik, J.; Gupta, A. Beyond skip connections: Top-down modulation for object detection. *arXiv* **2016**, arXiv:1612.06851.
28. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
29. Erhan, D.; Szegedy, C.; Toshev, A.; Anguelov, D. Scalable object detection using deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2147–2154.
30. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. *arXiv* **2019**, arXiv:1904.01355.
31. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
32. 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 October2017; pp. 2980–2988.
33. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
34. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9308–9316.
35. Xu, L.; Chen, Q. Remote-Sensing Image Usability Assessment Based on ResNet by Combining Edge and Texture Maps. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1825–1834. [CrossRef]
36. Yang, L.; Song, Q.; Wu, Y.; Hu, M. Attention Inspiring Receptive-Fields Network for Learning Invariant Representations. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 1744–1755. [CrossRef] [PubMed]
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
38. Ren, S.; He, K.; Girshick, R.; Zhang, X.; Sun, J. Object Detection Networks on Convolutional Feature Maps. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1476–1481. [CrossRef] [PubMed]
39. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995.
40. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Detnet: A backbone network for object detection. *arXiv* **2018**, arXiv:1804.06215.
41. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. DetNet: Design Backbone for Object Detection. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 339–354.
42. Samuel, N.; Diskin, T.; Wiesel, A. Learning to Detect. *IEEE Trans. Signal Process.* **2019**, *67*, 2554–2564. [CrossRef]
43. Tian, Z.; Wang, W.; Zhan, R.; He, Z.; Zhang, J.; Zhuang, Z. Cascaded Detection Framework Based on a Novel Backbone Network and Feature Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3480–3491. [CrossRef]
44. 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 October 2018; pp. 370–386.
45. Xie, X.; Yang, W.; Cao, G.; Yang, J.; Shi, G. The Collected XDUAV Dataset. Available online:. https://share.weiyun.com/8rAu3kqr (accessed on 10 September 2018).
46. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Zou, H. Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3652–3664. [CrossRef]

47. Sommer, L.W.; Schuchert, T.; Beyerer, J. Deep learning based multi-category object detection in aerial images. In Proceedings of the Automatic Target Recognition XXVII, Anaheim, CA, USA, 10–11 April 2017; Volume 10202, p. 1020209.
48. 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 (ICCV) Workshops, Seoul, Korea, 27–28 October 2019; pp. 118–126.
49. 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 (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
50. 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.
51. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
52. Majid Azimi, S. ShuffleDet: Real-Time Vehicle Detection Network in On-board Embedded UAV Imagery. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; pp. 88–99.
53. Huang, L.; Yang, Y.; Deng, Y.; Yu, Y. Densebox: Unifying landmark localization with end to end object detection. *arXiv* **2015**, arXiv:1509.04874.
54. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM international conference on Multimedia, Vancouver, BC, Canada, 19–24 October 2016; pp. 516–520.
55. Law, H.; Teng, Y.; Russakovsky, O.; Deng, J. CornerNet-Lite: Efficient Keypoint Based Object Detection. *arXiv* **2019**, arXiv:1904.08900.
56. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
57. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
58. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 850–859.
59. Chen, S.; Li, J.; Yao, C.; Hou, W.; Qin, S.; Jin, W.; Tang, X. DuBox: No-Prior Box Objection Detection via Residual Dual Scale Detectors. *arXiv* **2019**, arXiv:1904.06883.
60. Liu, W.; Liao, S.; Ren, W.; Hu, W.; Yu, Y. High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5187–5196.
61. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 840–849.
62. 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.
63. Newell, A.; Huang, Z.; Deng, J. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv* **2017**, arXiv:1611.05424.
64. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Shi, J. FoveaBox: Beyond Anchor-based Object Detector. *arXiv* **2019**, arXiv:1904.03797.
65. Chen, C.; Zhang, Y.; Lv, Q.; Wei, S.; Wang, X.; Sun, X.; Dong, J. RRNet: A Hybrid Detector for Object Detection in Drone-Captured Images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27–28 October 2019; pp. 100–108.
66. Cai, Y.; Du, D.; Zhang, L.; Wen, L.; Wang, W.; Wu, Y.; Lyu, S. Guided Attention Network for Object Detection and Counting on Drones. *arXiv* **2019**, arXiv:1909.11307.
67. Hu, P.; Ramanan, D. Finding tiny faces. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1522–1530.

68. Woo, S.; Hwang, S.; Kweon, I.S. StairNet: Top-Down Semantic Aggregation for Accurate One Shot Detection. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1093–1102.
69. Kong, T.; Sun, F.; Yao, A.; Liu, H.; Lu, M.; Chen, Y. Ron: Reverse connection with objectness prior networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5936–5944.
70. Kong, T.; Sun, F.; Tan, C.; Liu, H.; Huang, W. Deep feature pyramid reconfiguration for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 169–185.
71. Kong, T.; Yao, A.; Chen, Y.; Sun, F. Hypernet: Towards accurate region proposal generation and joint object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA , 26 June–1 July 2016; pp. 845–853.
72. Wang, H.; Wang, Z.; Jia, M.; Li, A.; Feng, T.; Zhang, W.; Jiao, L. Spatial Attention for Multi-Scale Feature Refinement for Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27–28 October 2019; pp. 64–72.
73. Li, J.; Wang, R.; Ding, J. Tiny Vehicle Detection from UAV Imagery. In *Image and Graphics Technologies and Applications*; Springer: Berlin, Germany, 2019.
74. Zhang, P.; Zhong, Y.; Li, X. SlimYOLOv3: Narrower, Faster and Better for Real-Time UAV Applications. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27–28 October 2019; pp. 37–45.
75. Ammar, A.; Koubaa, A.; Ahmed, M.; Saad, A. Aerial Images Processing for Car Detection using Convolutional Neural Networks: Comparison between Faster R-CNN and YoloV3. *arXiv* **2019**, arXiv:1910.07234.
76. Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector with Spatial Context Analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1758–1770. [CrossRef]
77. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
78. Liu, S.; Huang, D.; Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
79. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
80. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 658–666.
81. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F.F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 22–24 June 2009; pp. 248–255.
82. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.