remote sensing

MDPI

Article

Object Detection in UAV Images via Global Density Fused Convolutional Network

Ruiqian Zhang [1], Zhenfeng Shao [2,*], Xiao Huang [3], Jiaming Wang [2] and Deren Li [2]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; zhangruiqian@whu.edu.cn
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; wjmecho@whu.edu.cn (J.W.); drli@whu.edu.cn (D.L.)
[3] Department of Geosciences, University of Arkansas, Fayetteville, AR 72701, USA; xh010@uark.edu
* Correspondence: shaozhenfeng@whu.edu.cn

Received: 22 August 2020; Accepted: 21 September 2020; Published: 24 September 2020

Abstract: Object detection in Unmanned Aerial Vehicle (UAV) images plays fundamental roles in a wide variety of applications. As UAVs are maneuverable with high speed, multiple viewpoints, and varying altitudes, objects in UAV images are distributed with great heterogeneity, varying in size, with high density, bringing great difficulty to object detection using existing algorithms. To address the above issues, we propose a novel global density fused convolutional network (GDF-Net) optimized for object detection in UAV images. We test the effectiveness and robustness of the proposed GDF-Nets on the VisDrone dataset and the UAVDT dataset. The designed GDF-Net consists of a Backbone Network, a Global Density Model (GDM), and an Object Detection Network. Specifically, GDM refines density features via the application of dilated convolutional networks, aiming to deliver larger reception fields and to generate global density fused features. Compared with base networks, the addition of GDM improves the model performance in both recall and precision. We also find that the designed GDM facilitates the detection of objects in congested scenes with high distribution density. The presented GDF-Net framework can be instantiated to not only the base networks selected in this study but also other popular object detection models.

Keywords: object detection; UAV images; global density model; global density fused convolutional network

1. Introduction

Unmanned Aerial Vehicle (UAV) is a new and prominent remote sensing platform operated by radio remote control equipment or programming, which benefits a wide range of practical applications that include environmental monitoring [1–5], abnormal target tracking [6,7] and animal protection [8–10]. The rapid development in UAV techniques and applications has fostered wide attention in the object detection domain. In this paper, we focus on object detection in UAV images [11,12] that aims to identify and localize objects of interest from UAV images, serving as a basic and significant algorithm in numerous UAV applications.

In order to detect objects in UAV images, early algorithms adopt background extraction and selected feature extraction approaches [6,13–15]. Despite the effectiveness of these methods, they highly depend on the descriptive method of features and the perspective of images, which not only consume plenty of manpower and computation but also reduce the capability of the model in transferring on different datasets [8,16]. Over recent years, deep learning has become one of the most cutting-edge technologies in both computer vision and remote sensing communities [17–21]. Deep convolutional neural networks (DCNNs), an important network model in deep learning, brings significant progress

and achieves state-of-the-art performance in image analysis related fields. In object detection tasks, due to unprecedented success in deep learning based algorithms in natural scene images (such as the images in MS COCO [22], PASCAL [23] and ImageNet [24]), many researches adopt deep learning based algorithms in natural scene images to detect objects in UAV images [7,9,16,25]. However, the major difference between natural scene images and UAV images is that UAV images are often with varying scales, perspectives, and appearances, due to the fact that UAVs are maneuverable with high speed, multiple viewpoints, and altitudes [26–28]. In addition, unlike generic natural scenes with large individual objects, UAV images often contain a large number of small objects, leading to great challenges for object detection in UAV images using existing approaches [28,29].

To address the aforementioned challenges, we propose a global density fused convolutional network (termed as GDF-Net) that is able to cascade global features to facilitate object distribution learning, to detect objects in UAV images. The proposed method introduces congested scene analysis for dense object distribution learning, which is inspired by the methods in crowd counting tasks [30,31]. Compared with the existing networks, the proposed method improves the performance of congested scene object detection results, benefiting from object density features by dilated convolutional networks [30,32] and feature refinement. As shown in Figure 1, the architecture of the proposed GDF-Net consists of a Backbone Network, a Global Density Model (GDM), and an Object Detection Network. The innovative GDM fuses multiple features to a global density fused features using multi-level features from a Feature Pyramid Network (FPN) as inputs. Additionally, the GDM consists of a series of dilated convolutional networks where dilated kernels are applied to deliver larger reception fields of whole features from the backbone network. The generated global density fused features are integrated with the original features and promote feature alignment among objects distributed in congested scenes in UAV images, which further improve the performance of object detection in UAV images.

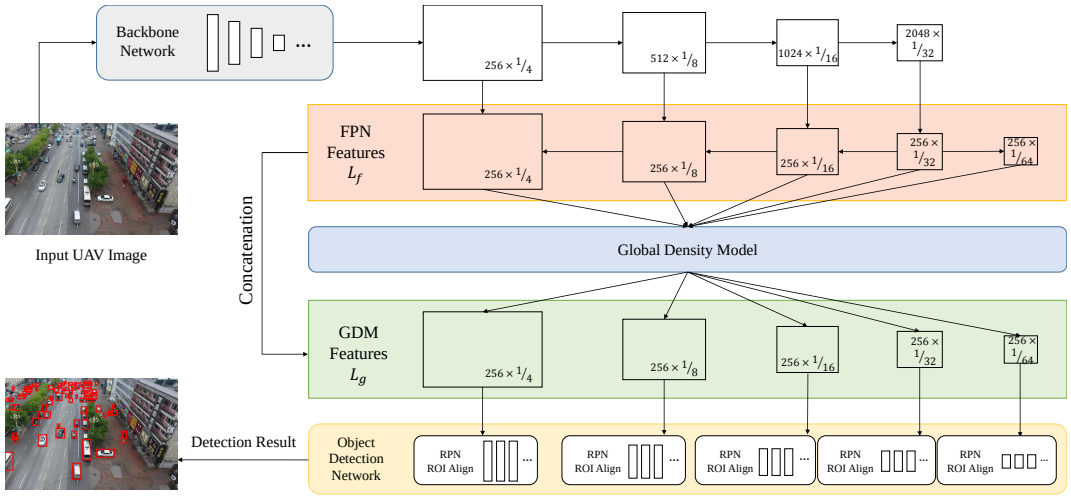


Figure 1. An overview of the proposed global density fused (GDF)-Net architecture. The GDF-Net consists of the Backbone Network, Global Density Model, and Object Detection Network, where the Backbone Network extracts pyramid features with typical networks, the Global Density Model integrates object distribution information into pyramid features, and the Object Detection Network locates and categorizes objects from UAV images.

We evaluate the proposed GDF-Net framework on two public UAV benchmark datasets: VisDrone [33] dataset and UAVDT [34]. As the proposed GDF-Net can be instantiated to existing detection algorithms, we perform several experiments of the GDF-Net instantiated on Faster R-CNN [35], Cascade R-CNN [36], Free Anchor [37] and Grid R-CNN [38]. To highlight the advantages of GDF-Net, we add GDM to the widely recognized algorithms above (the original algorithms are

called base networks) to detect objects in UAV images. The experimental results demonstrate that the proposed component improves the performance of object detection in UAV images.

The contributions of this work are summarized as follows:

- We propose a novel global density fused convolutional network (GDF-Net) for object detection in UAV images, which cascades a novel Global Density Model to base networks. Via the application of GDM, the proposed GDF-Net achieves a distribution learning that integrates global patterns from the input image with features extracted by existing object detection networks.
- We introduce a novel Global Density Model into the base networks to improve the performance of object detection in UAV images. GDM applies dilated convolutional networks to deliver large reception fields, facilitating the learning of global patterns in targets.
- The proposed GDF-Net can be instantiated to existing detection algorithms, and we demonstrate the effectiveness and robustness of GDF-Net on two popular UAV object detection datasets: VisDrone [33] dataset and UAVDT [34].

The remainder of the paper is organized as follows. Section 2 includes a brief review of the literature for object detection in UAV images. Section 3 presents the GDF-Net architecture and details the individual components within the framework. Section 4 describes the experimental settings, results, as well as the sensitivity analysis. Section 5 presents a discussion on the limitations and future directions, followed by conclusions in Section 6.

2. Related Work

Aiming to identify an object category, object detection has always been a hot topic in the computer vision domain. To detect objects in UAV images, early algorithms leverage the extraction of selected features and background information. Researchers in [6,13,14], detect objects in UAV images by generating a saliency map computed from the image background. Kalantar et al. [39] conduct object detection based on region adjacency graphs of visual appearance and geometric properties to facilitate background extraction from objects. Portmann et al. [15] detect pedestrians in UAV images using techniques in background subtraction and HoG feature extraction. Although these methods have been proved effective in terms of detection accuracy, they largely rely on the descriptive method of features and perspectives of images [8,16]. In addition, it is difficult for these methods to extract overlapping objects that commonly appear in congested regions, which largely reduces the generalization capability of these models in transferring among different tasks.

The recent development in object detection via deep learning methods has achieved unprecedented success [22]. These approaches can be generally divided into two categories: two-stage approach and one-stage approach. The two-stage algorithms detect objects based on both a region proposal network and object regression network (such as Faster R-CNN [35], Cascade R-CNN [36], Libra R-CNN [40], HyperNet [41], MS-CNN [42], CRAFT [43], FPN [44], etc.), while one-stage approaches focus on regression or classification networks without region proposals (such as YOLOv1-4 [45–48], SSD [49], G-CNN [50], DSSD [51], DSOD [52]). Compared with two-stage algorithms, one-stage methods are generally computationally efficient but with relatively lower accuracy.

The success of deep learning in identifying objects from natural images allows researches to adopt similar approaches for UAV images. For example, Wang et al. [25] experiment on numerous popular convolutional neural networks, such as SSD, Faster R-CNN, and RetinaNet, in natural images. Scholars in [53–55] develop various types of object detectors using enhanced deep convolutional neural networks based on SSD. Methodologies in [56,57] are designed based on improved YOLOv2 or YOLOv3 object detection algorithms, respectively. Different from early algorithms by selected features and background information, the deep learning based approaches automate object detection in UAV images and are transferable to different datasets. However, these object detection algorithms are designed on standard natural imagery and have not been optimized for detecting objects in UAV images. To improve the detection specifically in UAV images, scholars in [26,27] integrate object

detection and depth prediction for images obtained from micro UAVs. However, the inputs of stereo images greatly limit its potential application, as stereo images are often difficult to acquire.

3. Methodology

In this study, we propose a global density fused convolutional network (GDF-Net) for object detection in UAV images, which cascades a novel Global Density Model (GDM) to a base network, aiming to promote the object distribution learning. The GDM uses pyramid features from FPN and fuses multiple features to global density fused features. The proposed GDF-Net promotes feature alignment among objects distributed in congested scenes, which further improves the performance of object detection in UAV images. In the remainder of this section, we formulate proposed GDF-Net and detail the structures of the Backbone Network, the GDM, and the Object Detection Network.

3.1. Approach Overview

Let us consider a typical object detection problem over an input space $\mathcal{X}$ and a detection ground truth space $\mathcal{Y}$. The goal of an object detection algorithm is to learn the mapping $\mathcal{M}$ from $\mathcal{X}$ to $\mathcal{Y}$, i.e., $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$. In general, existing approaches based on deep learning algorithms firstly learn a deep feature $\mathcal{F}$ for the representative $\mathcal{X}$ and then obtain $\mathcal{Y}$ from $\mathcal{F}$ via object regression and a classification network. These algorithms can be abstracted using Equation (1), where $\mathcal{M}_1$ denotes mapping $\mathcal{M}$ from $\mathcal{X}$ to $\mathcal{Y}$. Given that UAVs are often with varying perspectives and flying altitudes, objects can be distributed in congested scenes, which increases the difficulty of object regression via only $\mathcal{F}$ from UAV images. To address this issue, our proposed GDF-Net applies a distribution learning that integrates the global features of an image with the features extracted from existing object detection networks. The GDF-Net introduces a Global Density Model (GDM), which learns a global feature $\mathcal{G}$ that describes object distribution in UAV images from input $\mathcal{X}$ and object density domain $\mathcal{H}$. Therefore, the mapping problem in our method can be defined as $\mathcal{M}_2$ in (2).

$$\mathcal{M}_1 = \mathcal{X} \xrightarrow{\mathcal{F}} \mathcal{Y} \tag{1}$$

$$\mathcal{M}_2 = \mathcal{X} \xrightarrow{\mathcal{G}} \mathcal{Y} = \mathcal{X} \xrightarrow{\mathcal{F},\mathcal{H}} \mathcal{Y} \tag{2}$$

An overview of our proposed GDF-Net approach is presented in Figure 1, where the GDM generates global density fused features for distribution learning using pyramid features. The Object Detection Network is leveraged to perform bounding box regression and target categorization using global density fused features as input.

3.2. Backbone Network

The goal of the Backbone Network is to generate high-dimensional features from input images by employing deep convolutional neural networks. In GDF-Net, Backbone Network is adopted from widely used feature extraction networks, such as ResNet-50 [58]. For an input image I, the Backbone Network obtains a feature collection L_b that contains deep features from the input image. As shown in Figure 1, we leverage a Feature Pyramid Network (FPN) [44] after the Backbone Network to detect multi-scale objects. If we define the Backbone Network and FPN respectively as $\mathcal{F}_b^1$ and $\mathcal{F}_b^2$, the FPN features L_f can be represented as:

$$L_f = \mathcal{F}_b(I) = \mathcal{F}_b^2(\mathcal{F}_b^1(I)) \tag{3}$$

where $L_f = \{L_f^1, L_f^2, ..., L_f^5\}$ are pyramid features from FPN (shown in the orange rectangular in Figure 1), enabling multi-scale object detection in UAV images. FPN takes a light top-down and bottom-up pathway with lateral connections to transform multi-level features to integrated pyramid features. Note that all features in L_f have 256 channels with different scales of feature height and width.

3.3. Global Density Model (GDM)

We design the Global Density Model with an aim to learn the global distribution of the targets. Specifically, we employ dilated convolutional networks, a technique to enlarge receptive fields and extract deeper features without losing resolutions [30], to obtain global density features. The architecture of GDM is inspired by the methods in crowd counting tasks where objects distributed in congested scenes usually create challenges for the counting algorithms. To solve this problem, scholars in [30,31] design dilated convolutional networks on deep features extracted from input images and produce density maps of images for crowd distribution analysis. Similarly, we believe that object detection from UAV images can also benefit from the learned object distribution, especially in congested scenes.

Compared with the approaches in crowd counting tasks that regress the counting number of objects from each image [30,31], object detection networks [35,36,38] focus on both object locations and categories determined by the multi-scale features extracted from deep convolutional neural networks. As stated in Section 3.2, pyramid features from FPN serve as input to GDM. In order to adopt distribution learning on integrated information from all pyramid features (L_f in Figure 1), our designed GDM is able to generate an integrated feature L_c from each input feature L_f (similar to the feature integrating operation in [31]). Taking the integrated feature L_c as input, GDM then produces a refined density feature L_d via dilated convolutional networks, for the purpose of enlarging receptive fields for distribution learning. In order to integrate density feature L_d and pyramid features L_f, the GDM architecture further scatters the refined density feature L_d to multi-level features L_g.

For detailed architecture, the GDM employs FPN features L_f (shown in the orange rectangle in Figure 1) and further acquires cascaded multiple features L_g (shown in the green rectangle in Figure 1) with global reception fields. The output features of GDM are the concatenation of input pyramid features and contain the information regarding the global density of targets. Assume $\mathcal{F}_g$ is the function of GDM, L_g can be obtained using the following formula:

$$L_g = \mathcal{F}_g(L_f) \tag{4}$$

Figure 2 illustrates a detailed structure of the proposed GDM. The process of designed GDM can be divided into the following three steps. We first generalize multiple FPN features L_f to a refined feature L_c, as the object distribution information requires features with an integrated perspective of the input image. After that, we generate a global feature L_d by transfering refined feature L_c to a density domain using dilated convolutional networks. Finally, through a residual path, refined density feature L_d is scattered to multi-level features L_g which are adopted as input for the Object Detection Network in GDF-Net. Specifically, an integrated multi-level features L_c can be presented as:

$$L_c = Avg(\sum_{i=1}^{5}(R_1(L_f^i, L_f^l))) \tag{5}$$

where parameter l, as the size of the l layer feature L_f^l, is defined to specify the size of the refined feature L_c, $Avg(\cdot)$ denotes average operation, and R_1 represents the resizing operation. Here, R_1 differs, given different layers within the feature L_f^i. If $i < l$, R_1 denotes a max pooling function P. Otherwise, R_1 denotes a resizing function U via the nearest interpolation. $A \xrightarrow{s} B$ means the operation with feature A according to the size of feature B. R_1 can be defined as:

$$R_1(L_f^i, L_f^l) = \begin{cases} P(L_f^i \xrightarrow{s} L_f^l), & i < l \\ U(L_f^i \xrightarrow{s} L_f^l), & i \geqslant l \end{cases} \tag{6}$$

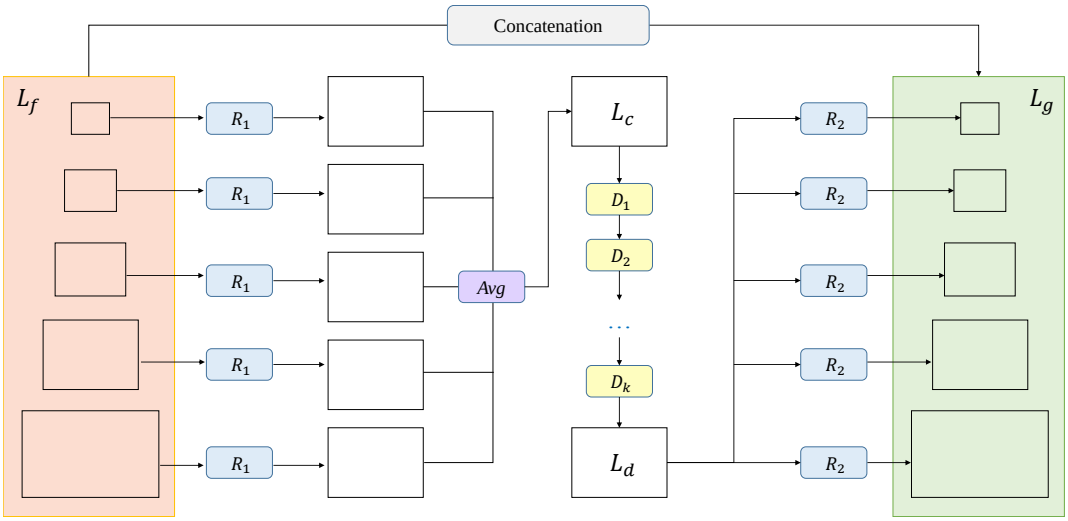


Figure 2. A detailed structure of the Global Density Model. In GDM, the FPN features L_f are fused to a refined feature L_c via a resizing operation R_1 shown with blue color and an average function Avg in the purple rectangle. The global feature L_d is obtained from dilated convolutional networks D and eventually scattered to multi-levels L_g by a residual path.

Aiming to enlarge receptive fields and consequently to improve detection accuracy, dilated convolutional layers have been proved rather efficient by multiple tasks [30,32,59]. Dilated convolutions, also called as atrous convolutions, introduce a new parameter, i.e., the dilation rate, aiming to set the number of interval pixel for convolution kernels [59]. The dilation rate defines the size of the reception field in GDM, which is crucial to the capacity of models in learning global features. The designed GDM adopts dilated convolutional layers to extract global distribution information from the refined feature L_c, further transferring it to L_d. Since L_d is refined from integrated features of the whole image, it represents the density domain with global information in our GDF-Net. The process in the generation of L_d can be represented as:

$$L_d = L_c \odot D_1(r) \odot D_2(r) \odot \cdot D_k(r) \tag{7}$$

where $\odot$ denotes the operation of convolution and $\{D_1, D_2, ...D_k\}$ denotes the dilated convolutional operation. Here, we set parameter r as the dilation rate for these dilated convolutional functions. If $r = 1$, a dilated convolution is the same as a normal convolution.

In order to apply the global density feature L_d in multiple scales, we employ a residual path to scatter L_d to multi-level features L_g using the following formula:

$$R_2(L_d, L_f^j) = \begin{cases} U(L_d \xrightarrow{s} L_f^j), & j < l \\ P(L_d \xrightarrow{s} L_f^j), & j \geqslant l \end{cases} \tag{8}$$

$$L_g^j = R_2(L_d, L_f^j) \oplus L_f^j \tag{9}$$

where $\oplus$ indicates the element-wise sum operator, which introduces a residual path between L_d and L_f.

3.4. Object Detection Network

To learn the explicit mapping from extracted global density fused features $\mathcal{G}$ to ground truth space $\mathcal{Y}$, we apply Object Detection Network, which regresses the position of each object by a regressor and classifies its category by classifier. The Object Detection Network takes L_g as input and generates

location p_{loc} and category p_{cls} of each object prediction n. Compared with the ground-truthing value $\{g_{loc}, g_{cls}\}$, the loss function GDF-Net during the training process is calculated as:

$$L = \Delta_1(p_{loc}, g_{loc}) + \lambda \cdot \Delta_2(p_{cls}, g_{cls}) \tag{10}$$

where $\Delta_1(\cdot)$ and $\Delta_2(\cdot)$ define the rules in location and category offset calculation, respectively. λ denotes the weight between the loss of regressor and classifier. In our proposed GDF-Net, the architecture of the Object Detection Network can be instantiated into different existing algorithms, such as Faster R-CNN [35], YOLO [45], and Cascade R-CNN [36]. Here, we apply it to numerous two-stage detection networks in Figure 1, which adopts Region Proposal Network (RPN) and ROI Align structures [35]. $\Delta_1(\cdot)$, $\Delta_2(\cdot)$ denote the error calculation methods that vary depending on the choice of the algorithms (e.g., $\Delta_1(\cdot)$ denotes SmoothL1 loss and $\Delta_2(\cdot)$ means Cross Entropy Loss in Faster R-CNN [35]).

4. Experimental Results and Analysis

We conduct a series of experiments to evaluate the performance of the proposed GDF-Net. In this section, we describe the benchmark datasets, evaluation metrics, and implementation details used in our training and testing experiments. We compare our method with several popular object detection approaches quantitatively and qualitatively to shed light on the advantage of the proposed GDF-Net framework. In addition, we further analyze the effects of the setting of dilation rate r on the performance of GDF-Net.

4.1. Experimental Setup

4.1.1. Datasets

Two challenging UAV benchmark datasets were evaluated by the proposed GDF-Net, i.e., VisDrone dataset [33] and UAVDT dataset [34]. The selected two datasets well simulate various scenarios of UAV object detection, as they contain objects obtained from different sensors, with varying weather conditions, perspectives, flying altitudes, camera views, and occlusions. In our experiments, we focused on presenting the improvement in performance when the proposed GDF-Net framework is added to the selected base networks. Specifically, the original VisDrone dataset [33] consists of a total of 400 video clips. Our training set, validation set, and testing set respectively contain a total of 6471, 548, and 1610 images. The VisDrone dataset labels humans and vehicles in daily life in ten categories, i.e., pedestrian, person, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. Another benchmark dataset is the UAVDT dataset [34], which consists of UAV imagery with vehicles in three categories (car, truck, and bus) selected from videos (about 10-h long). In the UAVDT dataset, we derived 11,915 images for training and 16,580 images for testing.

4.1.2. Evaluation Metrics

To evaluate the performance of GDF-Net, we employed precision metrics and recall metrics, defined in [22]. The precision metrics, including AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L, were calculated as the ratio of the average correctly predicted positive observations to the total number of predicted positive observations. The recall metrics, including AR_1, AR_{10}, AR_{100}, AR_S, AR_M and AR_L, were calculated as the ratio of the correctly predicted positive observations to all observations. Here, the AP, AP_S, AP_M, AP_L, and all recall metrics use ten intersections over union (IOU) values ($[0.50 : 0.05 : 0.95]$) as IOU thresholds to calculate average precision and recall results. The AP50 and AP75 evaluate results using IOU thresholds as 0.5 and 0.75, respectively. Moreover, $\{S, M, L\}$ in these indexes represent the average precision at different scales. The recall index APnum indicates the maximum recall given num detection per image for $num = \{1, 10, 100\}$. A detailed definition of these

metrics can be found in [22]. These metrics are widely used in the existing object detection literature, and they evaluate the performance of proposed GDF-Net in a comprehensive manner.

4.1.3. Implementation Details

The proposed GDF-Net framework was implemented with the PyTorch framework and was run at a desktop equipped with an Intel(R) Core(TM) i7-9800X CPU @ 3.80 GHz, two NVIDIA Geforce RTX 2080ti GPUs with 11 G memory each. All experiments were conducted on Ubuntu 16.04 system with two parallel GPUs. The whole program is implemented based on the publicly available Open MMLab Detection [60] framework on the PyTorch platform. We initiate the Backbone Network parameters in the ResNet50 [58] model (pre-trained on ImageNet). Other parameters in GDM and Object Detection Network are randomly initialized.

We focused on evaluating the performance of the proposed GDF-Net framework when it is attached to existing popular base networks. All experiments were conducted following the default parameter settings in the base networks and without data augmentation. In light of the different scales of images from the VisDrone dataset, we resize them to 1200×675 pixels during the training process. Images from the UAVDT dataset were fed to GDF-Net with their original size, i.e., 1024×540. Moreover, we chose Stochastic Gradient Descent (SGD) as the optimizer [61] and set the momentum as 0.9, weight decay as 10^{-4}, the initial learning rate as 0.02. The experiments using Grid R-CNN were trained with 25 epochs [38], and all other experiments were trained with 12 epochs. In our experiments, we empirically set the $l = 2$ and $\lambda = 1$ in Section 3.4. The dilation rate r in Section 3.3 was set to 2 based on experiments in Section 4.2.3. In addition, we set k as 6, and the channels of $\{D_1, D_2, ..., D_k\}$ as $[512, 512, 512, 256, 128, 64]$, following the experiments in [30].

4.2. Evaluation of Gdf-Net

4.2.1. Quantitative Evaluation

We evaluated the performance of the proposed GDF-Net framework with state-of-the-art object detection methods by adding the designed GDM to these existing networks (the base networks). These base networks include Faster R-CNN [35], Cascade R-CNN [36], Free Anchor [37], and Grid R-CNN [38]. We term them Faster GDF, Cascade GDF, Free Anchor GDF, and Grid GDF, when GDM is respectively added to Faster R-CNN, Cascade R-CNN, Free Anchor, and Grid R-CNN. All parameters remain unchanged in each set of comparisons. The quantitative evaluations in the VisDrone dataset and UAVDT dataset can be found respectively in Tables 1 and 2.

Table 1. Comparisons of detection performance on the VisDrone dataset.

Method	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L
Faster R-CNN	31.0	17.5	17.2	8.0	26.9	34.9	7.8	23.5	28.2	16.5	42.8	**50.3**
Faster GDF (ours)	**31.8**	**17.9**	**17.7**	**8.2**	**27.7**	**35.8**	**7.9**	**23.8**	**28.8**	**17.0**	**43.7**	49.7
Cascade R-CNN	31.1	19.3	18.3	8.5	28.3	36.3	8.2	23.8	28.4	16.8	42.7	50.2
Cascade GDF (ours)	**31.7**	**19.4**	**18.7**	**8.7**	**28.7**	**38.7**	**8.4**	**24.2**	**28.8**	**17.0**	**43.5**	**52.4**
Free Anchor	27.9	15.8	15.6	7.1	23.7	28.8	7.0	22.1	29.9	**18.8**	**41.9**	55.1
Free Anchor GDF (ours)	**28.5**	**16.0**	**15.9**	**7.2**	**23.7**	**33.5**	**7.1**	**22.1**	**29.9**	18.7	41.8	**56.0**
Grid R-CNN	30.4	18.9	17.9	8.3	27.8	35.7	8.1	23.9	28.5	17.2	42.8	49.7
Grid GDF (ours)	**30.8**	**19.2**	**18.2**	**8.6**	**27.9**	**37.9**	**8.1**	**24.1**	**28.7**	**17.3**	**43.0**	**52.8**

Note: better performances are highlighted in bold.

Table 2. Comparisons of detection performance on the UAVDT dataset.

Method	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L
Faster R-CNN	25.8	14.7	14.4	9.1	25.2	21.8	13.2	22.5	27.3	14.8	42.9	40.6
Faster GDF (ours)	**27.5**	**16.7**	**15.6**	**9.6**	**27.1**	**25.4**	**13.4**	**23.6**	**28.2**	**15.4**	**44.2**	**42.3**
Cascade R-CNN	25.3	16.0	14.8	**9.7**	25.2	**28.0**	12.3	22.0	26.6	**15.6**	40.7	**45.0**
Cascade GDF (ours)	**26.0**	**16.2**	**15.0**	9.4	**26.1**	23.1	**12.4**	**22.4**	**27.1**	14.8	**42.4**	44.4
Free Anchor	27.9	16.0	15.6	**9.9**	26.5	22.9	13.9	24.5	29.4	**16.3**	45.4	**40.3**
Free Anchor GDF (ours)	**27.9**	**16.3**	**15.7**	9.5	**27.5**	**25.0**	**14.3**	**24.8**	**29.4**	15.8	**46.4**	39.4
Grid R-CNN	24.5	15.7	14.4	8.7	25.4	**25.5**	12.2	22.5	27.0	15.2	42.1	40.6
Grid GDF (ours)	**26.1**	**17.0**	**15.4**	**8.9**	**27.3**	24.4	**13.2**	**23.1**	**27.6**	**15.2**	**43.3**	**40.9**

Note: better performances are highlighted in bold.

In general, our method achieves the best performance on both datasets, in almost all precision and recall metrics (up to 1.2% gains in AP and 0.9% gains in AR_{100}). The experiments on multiple base algorithms demonstrate that our proposed GDF-Net network improves both precision and recall compared to the base algorithms alone, suggesting its robustness and compatibility. It is worth noting, however, that a few base networks underperform when designed GDM is added to them, especially under the evaluation using AP_L. Compared with the AP_S and AP_M, AP_L has seen the most improvement in various experiments. The results reveal that the GDF-Net well-performs in terms of learning global distribution patterns, retaining the balance among objects with scale diversity. However, improving the detection performance of objects in varying scales simultaneously still remains a challenge, and the detection accuracy of large objects is considerably higher than that of small objects (up to 4.7% gains in AP_L and 0.3% gains in AP_S in Table 1). Therefore, given the trade-off process of detecting large objects and small objects in UAV images via a certain model, the improvement of precision of small targets has a greater significance in overall detection evaluation AP.

To highlight the utility of the proposed method, we present the accuracy, complexity, and speed between baseline models and baseline models coupled with GDF-Net on the UAVDT dataset (Table 3). The Params and FLOPs respectively denote the number of parameters and speed of performing multiply-adds [62], while the speed measures the processing speed of scenes using frame per second (fps) as a unit [57]. For a fair comparison, results are measured on the same GPU with the same settings described in Section 4.1.3. We observe that, with the attachment of GDF-Net, the number of parameters has slightly increased, resulting in slightly reduced FLOPS and speed. However, considering the improvement of the general accuracy and the improvement of detection, especially in congested scenes (illustrated in Section 4.2.2), we believe the advantages of attaching GDF-Net outweigh the disadvantages.

Table 3. Accuracy, complexity and speed comparison on the UAVDT Dataset.

Experiment	AP_{50}	Params	FLOPs	Speed
Faster R-CNN	25.8	41.2 M	118.8 GMac	23.7 fps
Faster GDF (ours)	27.5	48.9 M	135.1 GMac	21.1 fps
Cascade R-CNN	25.3	69.0 M	146.6 GMac	17.6 fps
Cascade GDF (ours)	26.0	76.7 M	162.8 GMac	16.2 fps
Free Anchor	27.9	36.3 M	113.5 GMac	24.3 fps
Free Anchor GDF (ours)	27.9	44.0 M	117.6 GMac	22.6 fps
Grid R-CNN	24.5	64.3 M	241.4 GMac	19.4 fps
Grid GDF (ours)	26.1	72.0 M	257.6 GMac	17.9 fps

4.2.2. Qualitative Evaluation

We present the experimental results generated by the proposed GDF-Net in Figure 3 (VisDrone dataset) and Figure 4 (UAVDT dataset), where all objects are shown in green rectangles marked by their categories. These detection results are based on the Faster GDF-Net method trained on two benchmark datasets independently. From the results, we observe that the objects are successfully detected with great accuracy, despite the heterogeneity in backgrounds, perspectives, flying altitudes, scales, and appearances, demonstrating the great performance of the proposed GDF-Net in object detection from UAV images. However, we also find that GDF-Net fails to detect obscured objects or objects with a hazy background (see the last row in Figures 3 and 4), suggesting that further improvement is still needed. Furthermore, for the purpose of visualizing the comparison between our proposed networks and baselines, we randomly selected examples of detection results on Faster R-CNN and Faster GDF-Net on the VisDrone dataset (Figure 5). From the zoom-in views (red rectangles), we observe that the proposed GDF-Net is able to detect more objects in highly congested regions (see the comparison between Figure 5k,l). We also notice that GDF-Net produces less wrong detections in highly congested regions (see the comparison between Figure 5g,h), presumably due to the additional global density model (GDM).

Figure 3. Detection results of the proposed GDF-Net approach on the VisDrone dataset. Detections are labeled in green rectangles and marked with associated categories and confidence scores.

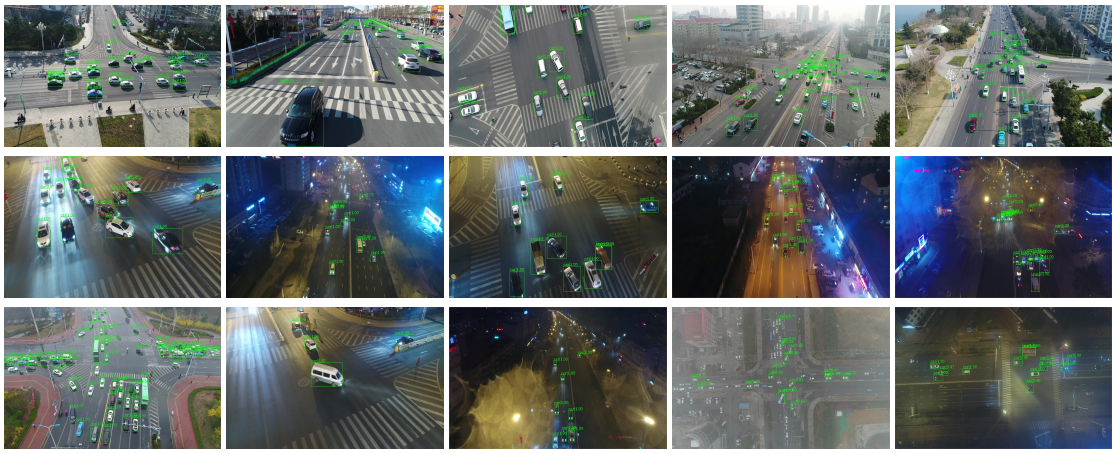
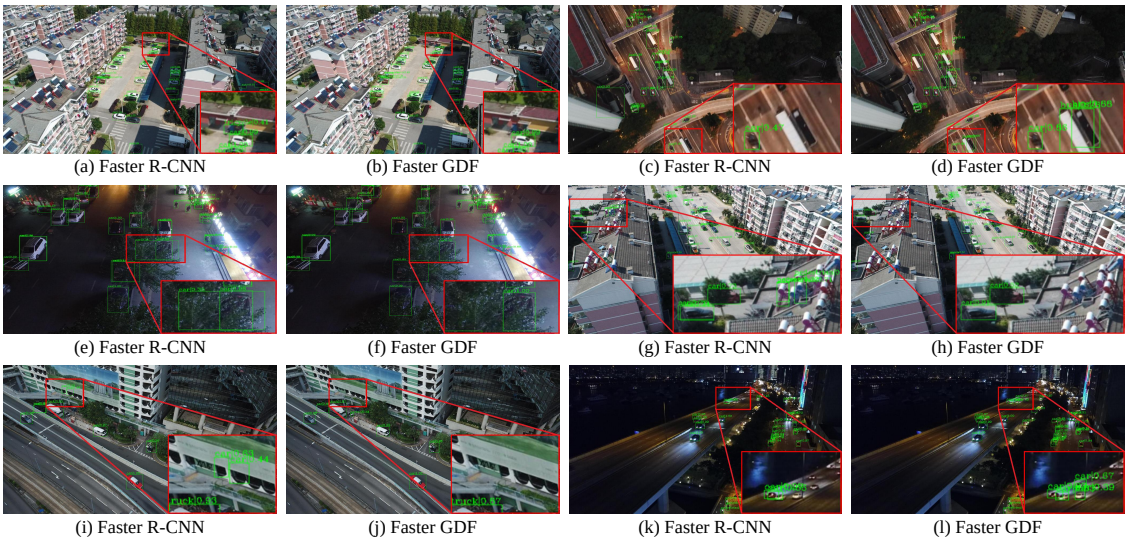


Figure 4. Detection results of the proposed GDF-Net approach on the UAVDT dataset. Detections are labeled with green rectangles and marked with associated categories and confidence scores.



(a) Faster R-CNN (b) Faster GDF (c) Faster R-CNN (d) Faster GDF

(e) Faster R-CNN (f) Faster GDF (g) Faster R-CNN (h) Faster GDF

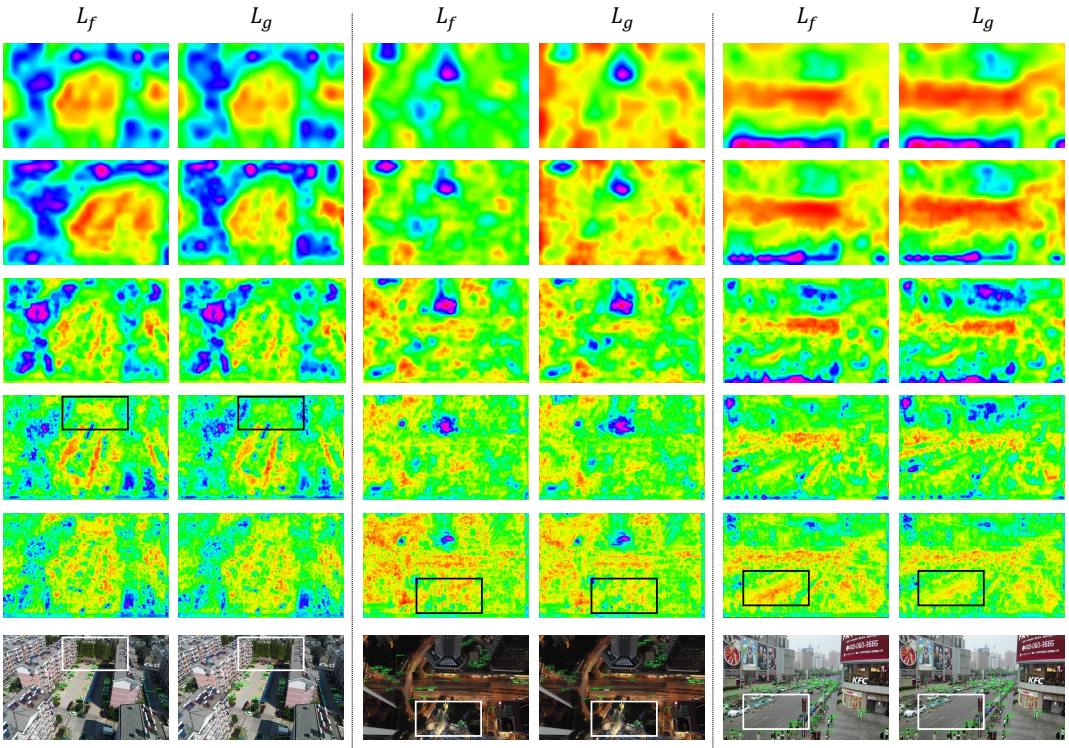
(i) Faster R-CNN (j) Faster GDF (k) Faster R-CNN (l) Faster GDF

Figure 5. Selected examples of detection results based on the Faster R-CNN and the proposed Faster GDF approach. Compared with results from Faster R-CNN, the proposed Faster GDF detects more objects in dense distribution with lower False Positives.

We further visualize L_f and L_g in base networks alone and base networks combined with designed GDM (Figure 6). As described in Section 3.3, $L_f = \{L_f^1, L_f^2, ..., L_f^5\}$ are pyramid features from FPN, serving as inputs to GDM, while $L_g = \{L_g^1, L_g^2, ..., L_g^5\}$ are generated by GDM with numerous operations that include gathering, refining, and scattering. Thus, the visualization of L_f and L_g explicitly reflects the functionality of GDM, which contributes to the explanation of the difference between base models and GDF-Nets. Images in row 1–5 and column 1, 3, 5 in Figure 6 are the visualizations of $\{L_f^1, L_f^2, ..., L_f^5\}$ while images in row 1–5 and column 2, 4, 6 are visualizations of the corresponding $\{L_g^1, L_g^2, ..., L_g^5\}$. All images are presented in HSV color space using channel maximum squeeze. The comparison between L_f and L_g illustrates that L_g is generally more accurate compared with the corresponding L_f, evidenced by the fact that L_g features show more consistency with input image than L_f (see regions marked by the black and white rectangles in Figure 6). From the detected objects (last row in Figure 6), the proposed GDF-Net achieves better performance, as a certain amount of objects are neglected by Faster R-CNN but successfully detected by Faster GDF (e.g., the white vehicle in the lower right corner).

4.2.3. Sensitivity Analysis

As the dilation rate, i.e., parameter r, defines the size of the reception field in GDM, which is crucial to the capacity of models in learning global features. To optimize r, we test different values of r (equal intervals from 1 to 3) on Faster GDF, Cascade GDF, and Grid GDF, which show high performance in Table 1. The experiments are conducted on the VisDrone dataset. As shown in Table 4, regardless of the setting of r, GDF-Nets generally outperform the corresponding base networks, evidenced by the higher values in both precision and recall. As r increases from 1 to 2, almost all evaluation indexes increase accordingly, suggesting that the performance of our network improves with a slightly larger receptive field that facilitates the extraction of global distribution features. However, from $r = 2$ to $r = 3$, model performance generally reduces, indicating that an excessive enlargement of r limits the improvement of the performance. We can conclude that $r = 2$ is the optimized value in this study, and it is important for the proposed GDF-Net to keep the balance between enlarged reception field with larger r and obtain a detailed structure with smaller r.



(a) Faster R-CNN (b) Faster GDF (c) Faster R-CNN (d) Faster GDF (e) Faster R-CNN (f) Faster GDF

Figure 6. Visualization of L_f and L_g features in the proposed GDF-Net. The images in row 1–5 show the visualization results of five layer L_f or L_g based on Faster GDF, respectively. The last row presents object detection results of three images (every two pictures are results with the same input image) from VisDrone dataset, where (**a**,**c**,**e**) are tested by Faster R-CNN, and (**b**,**d**,**f**) are detected from Faster GDF, respectively.

Table 4. Sensitivity analysis on the VisDrone dataset. We analyze the effects of r setting in Section 3.3 with multiple algorithms, and all experiments are performed with GDF-Net.

Method	Para	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L
Faster GDF	r = 1	31.7	17.8	17.7	8.2	27.5	35.2	7.8	23.6	28.5	16.8	43.4	50.5
Faster GDF	r = 2	**31.8**	**17.9**	**17.7**	**8.2**	**27.7**	**35.8**	**7.9**	**23.8**	**28.8**	**17.0**	**43.7**	**49.7**
Faster GDF	r = 3	31.5	**18.0**	17.6	8.0	27.6	34.8	7.9	23.7	28.7	16.8	43.7	49.4
Cascade GDF	r = 1	31.2	19.2	18.4	8.5	28.3	36.9	8.1	23.9	28.6	16.8	43.0	52.3
Cascade GDF	r = 2	**31.7**	19.4	**18.7**	**8.7**	28.7	**38.7**	**8.4**	**24.2**	**28.8**	**17.0**	43.5	**52.4**
Cascade GDF	r = 3	31.7	**19.8**	18.7	8.5	**29.1**	38.7	8.4	24.2	28.8	16.9	**43.8**	50.9
Grid GDF	r = 1	30.6	19.0	18.0	8.4	27.8	36.7	8.0	24.1	28.8	16.5	42.8	50.3
Grid GDF	r = 2	**30.8**	19.2	**18.2**	**8.6**	27.9	**37.9**	**8.1**	24.1	**28.7**	17.3	**43.3**	50.9
Grid GDF	r = 3	30.5	**19.3**	18.2	8.4	**28.0**	37.3	8.1	**24.2**	28.7	**17.5**	43.2	**51.4**

Note: better performances are highlighted in bold.

5. Limitations and Future Directions

Although the proposed GDF-Net brings promising results for object detection in UAV images, some notable issues remain and call for further research. Firstly, four popular algorithms are selected as base algorithms in our experiment. However, we acknowledge that algorithms that include HTC [63] and YOLOv4 [48] have become more popular recently. Thus, the potential of those methods in the proposed framework deserves further investigation.

Secondly, to fairly evaluate the performance of designed GDM on existing base networks and mitigate the impacts resulting from the difference in parameter settings, we set parameters (e.g., $\Delta_1(\cdot)$, $\Delta_2(\cdot)$ and l) according to the empirical values from the base algorithms, and include no data augmentation in all experiments. However, researches have shown that the parameter, loss settings, and data augmentation have a great impact on the performance of deep learning models [64–67]. Further research is needed to further optimize relevant parameters, experiment loss settings, and employ data augmentation methods in these UAV image datasets.

Thirdly, despite the fact that our proposed GDF-Net improves the performance compared to the base networks, it usually fails to detect occluded objects, especially in congested scenes and with a hazy background. Fortunately, numerous studies have been conducted to address this issue, most notable of which is by [68], who applied a novel aggregation loss function and a pooling method for occlusion detection, providing a great opportunity to identify partially obscured objects. In future studies, we plan to incorporate the aforementioned methods to our GDF-Net.

Lastly, dilated convolutional networks are applied to deliver larger reception fields and generate global density fused features. Although the great utility of dilated convolutional networks has been proved in this study, other emerging techniques, for example, attention mechanism [69] and Generative Adversarial Network (GAN) [70], have received growing attention. The potential of those methods in rendering large reception fields and how they can be incorporated in the proposed GDF-Net framework deserve further exploration.

6. Conclusions

Object detection in UAV imagery remains a challenging task, as UAVs are often maneuverable with high speed, multiple viewpoints, and varying altitudes, which leads to unique characteristics of UAV imagery that usually contain varying perspectives, scales, and occlusion. In addition, objects in UAV images are often distributed with heterogeneity, varying in size, with high density, causing great difficulty for object detection using existing algorithms that are not optimized for UAV images. In this paper, we propose a novel global density fused convolutional network (GDF-Net) specifically for object detection in UAV images. The proposed GDF-Nets consists of a Backbone Network, a Global Density Model (GDM), and an Object Detection Network. We test the effectiveness and robustness of the proposed GDF-Nets on the VisDrone dataset and UAVDT dataset. The novelty in GDM is that it refines density features via the application of dilated convolutional networks, aiming to deliver larger reception fields and to facilitate the generation of global density fused features. When comparing with the scenario where base networks are used independently, the addition of GDM improves the model performance in both recall and precision. We also find that the designed GDM facilitates the detection of objects in congested scenes with high distribution density. The presented GDF-Net framework can be instantiated to not only the base networks selected in this study but also other popular object detection models.

Author Contributions: Conceptualization, R.Z. and J.W.; methodology, software, and validation, R.Z. and J.W.; formal analysis and investigation, R.Z.; resources, Z.S.; writing–original draft preparation, R.Z. and X.H.; writing–review and editing, Z.S. and X.H.; supervision, X.H. and D.L.; project administration and funding acquisition, Z.S. and D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National key R and D plan on strategic international scientific and technological innovation cooperation special project: 2018YFB2100501; The National Natural Science Foundation of China: 61671332, 41771452, 51708426, 41890820 and 41771454; The Natural Science Fund of Hubei Province in China: 2018CFA007 and The Independent Research Projects of Wuhan University: 2042018kf0250.

Acknowledgments: We would like to thank the anonymous reviewers for their constructive and valuable suggestions on the earlier drafts of this manuscript.

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

References

1. Zhang, R.; Li, H.; Duan, K.; You, S.; Liu, K.; Wang, F.; Hu, Y. Automatic Detection of Earthquake-Damaged Buildings by Integrating UAV Oblique Photography and Infrared Thermal Imaging. *Remote Sens.* **2020**, *12*, 2621. [CrossRef]
2. Zhou, G.; Ambrosia, V.; Gasiewski, A.J.; Bland, G. Foreword to the special issue on unmanned airborne vehicle (UAV) sensing systems for earth observations. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 687–689. [CrossRef]
3. Hird, J.N.; Montaghi, A.; McDermid, G.J.; Kariyeva, J.; Moorman, B.J.; Nielsen, S.E.; McIntosh, A. Use of unmanned aerial vehicles for monitoring recovery of forest vegetation on petroleum well sites. *Remote Sens.* **2017**, *9*, 413. [CrossRef]
4. Shao, Z.; Li, C.; Li, D.; Altan, O.; Zhang, L.; Ding, L. An Accurate Matching Method for Projecting Vector Data into Surveillance Video to Monitor and Protect Cultivated Land. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 448. [CrossRef]
5. Li, W.; Fu, H.; Yu, L.; Cracknell, A. Deep Learning Based Oil Palm Tree Detection and Counting for High-Resolution Remote Sensing Images. *Remote Sens.* **2016**, *9*, 22. [CrossRef]
6. Li, Y.; Zhang, Y.; Yu, J.G.; Tan, Y.; Tian, J.; Ma, J. A novel spatio-temporal saliency approach for robust dim moving target detection from airborne infrared image sequences. *Inf. Sci.* **2016**, *369*, 548–563. [CrossRef]
7. Kapania, S.; Saini, D.; Goyal, S.; Thakur, N.; Jain, R.; Nagrath, P. Multi Object Tracking with UAVs using Deep SORT and YOLOv3 RetinaNet Detection Framework. In Proceedings of the 1st ACM Workshop on Autonomous and Intelligent Mobile Systems, Aviero, Portugal, 25–27 June 2020; pp. 1–6.
8. Benjamin, K.; Diego, M.; Devis, T. Detecting Mammals in UAV Images: Best Practices to address a substantially Imbalanced Dataset with Deep Learning. *Remote Sens. Environ.* **2018**, *216*, 139–153.
9. Kellenberger, B.; Volpi, M.; Tuia, D. Fast animal detection in UAV images using convolutional neural networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017.
10. 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]
11. Meng, L.; Peng, Z.; Zhou, J.; Zhang, J.; Lu, Z.; Baumann, A.; Du, Y. Real-Time Detection of Ground Objects Based on Unmanned Aerial Vehicle Remote Sensing with Deep Learning: Application in Excavator Detection for Pipeline Safety. *Remote Sens.* **2020**, *12*, 182. [CrossRef]
12. Qian, X.; Lin, S.; Cheng, G.; Yao, X.; Wang, W. Object Detection in Remote Sensing Images Based on Improved Bounding Box Regression and Multi-Level Features Fusion. *Remote Sens.* **2020**, *12*, 143. [CrossRef]
13. Wang, G. Vision-Based Real-Time Aerial Object Localization and Tracking for UAV Sensing System. *IEEE Access* **2017**, *5*, 23969–23978.
14. Cong, M.; Han, L.; Ding, M.; Xu, M.; Tao, Y. Salient man-made object detection based on saliency potential energy for unmanned aerial vehicles remote sensing image. *Geocarto Int.* **2019**, *34*, 1634–1647. [CrossRef]
15. Portmann, J.; Lynen, S.; Chli, M.; Siegwart, R. People detection and tracking from aerial thermal views. In Proceedings of the IEEE international conference on robotics and automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 1794–1800.
16. Bazi, Y.; Melgani, F. Convolutional SVM networks for object detection in UAV imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3107–3118. [CrossRef]
17. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
18. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
19. Papadomanolaki, M.; Vakalopoulou, M.; Karantzalos, K. A novel object-based deep learning framework for semantic segmentation of very high-resolution remote sensing data: Comparison with convolutional and fully convolutional networks. *Remote Sens.* **2019**, *11*, 684. [CrossRef]
20. Shao, Z.; Zhou, W.; Deng, X.; Zhang, M.; Cheng, Q. Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 318–328. [CrossRef]

21. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction From High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
22. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
23. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2012; pp. 1097–1105.
25. Wang, X.; Cheng, P.; Liu, X.; Uzochukwu, B. Fast and accurate, convolutional neural network based approach for object detection from UAV. In Proceedings of the IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society, Washington, DC, USA, 21–23 October 2018; pp. 3171–3175.
26. Aguilar, W.G.; Quisaguano, F.J.; Rodríguez, G.A.; Alvarez, L.G.; Limaico, A.; Sandoval, D.S. Convolutional neuronal networks based monocular object detection and depth perception for micro UAVs. In *International Conference on Intelligent Science and Big Data Engineering*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 401–410.
27. Carrio, A.; Vemprala, S.; Ripoll, A.; Saripalli, S.; Campoy, P. Drone detection using depth maps. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1034–1037.
28. Chen, C.; Zhang, Y.; Lv, Q.; Wei, S.; Dong, J. RRNet: A Hybrid Detector for Object Detection in Drone-Captured Images. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Seoul, Korea, 27–28 October 2019; pp. 100–108.
29. 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 Workshop, Seoul, Korea, 27–28 October 2019; pp. 118–126.
30. Li, Y.; Zhang, X.; Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1091–1100.
31. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 589–597.
32. Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 472–480.
33. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Nie, Q.; Cheng, H.; Liu, C.; Liu, X.; et al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
34. 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.
35. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
36. 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, 18–23 June 2018; pp. 6154–6162.
37. Zhang, X.; Wan, F.; Liu, C.; Ji, R.; Ye, Q. FreeAnchor: Learning to Match Anchors for Visual Object Detection. In *Neural Information Processing Systems (NIPS)*; NIPS: Grenada, Spain, 2019.
38. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid R-CNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
39. Kalantar, B.; Mansor, S.B.; Halin, A.A.; Shafri, H.Z.M.; Zand, M. Multiple Moving Object Detection from UAV Videos Using Trajectories of Matched Regional Adjacency Graphs. *IEEE Trans. Geosci. Remote Sens.* **2017**, 1–16. [CrossRef]

40. 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 Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.
41. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
42. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 354–370.
43. Yang, B.; Yan, J.; Lei, Z.; Li, S.Z. Craft objects from images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 6043–6051.
44. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
45. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
46. 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. 7263–7271.
47. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
48. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
49. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
50. Najibi, M.; Rastegari, M.; Davis, L.S. G-cnn: An iterative grid based object detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2369–2377.
51. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
52. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.G.; Chen, Y.; Xue, X. Dsod: Learning deeply supervised object detectors from scratch. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1919–1927.
53. Kyrkou, C.; Plastiras, G.; Theocharides, T.; Venieris, S.I.; Bouganis, C.S. DroNet: Efficient convolutional neural network detector for real-time UAV applications. In Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), Dresden, Germany, 19–23 March 2018; pp. 967–972.
54. Li, Y.; Dong, H.; Li, H.; Zhang, X.; Zhang, B.; Xiao, Z. Multi-block SSD Based Small Object Detection for UAV Railway Scene Surveillance. *Chin. J. Aeronaut.* **2020**. [CrossRef]
55. 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]
56. Tijtgat, N.; Ranst, W.V.; Volckaert, B.; Goedeme, T.; Turck, F.D. Embedded Real-Time Object Detection for a UAV Warning System. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Venice, Italy, 22–29 October 2017.
57. 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 Workshop, Seoul, Korea, 27–28 October 2019; pp. 37–45.
58. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
59. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
60. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.

61. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the COMPSTAT'2010, Paris, France, 22–27 August 2010.
62. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Seattle, WA, USA, 14 June 2020; pp. 10781–10790.
63. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4974–4983.
64. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **2010**, *9*, 249–256.
65. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
66. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5561–5569.
67. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
68. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 637–653.
69. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
70. Li, J.; Liang, X.; Wei, Y.; Xu, T.; Feng, J.; Yan, S. Perceptual generative adversarial networks for small object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1222–1230.