remote sensing

MDPI

Technical Note

Small-Sized Vehicle Detection in Remote Sensing Image Based on Keypoint Detection

Lijian Yu †, Xiyang Zhi *,†, Jianming Hu, Shikai Jiang, Wei Zhang and Wenbin Chen

Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China; 19B921013@stu.hit.edu.cn (L.Y.); hujianming@stu.hit.edu.cn (J.H.); 18B921009@stu.hit.edu.cn (S.J.); wzhang@hit.edu.cn (W.Z.); 18B921010@stu.hit.edu.cn (W.C.)

* Correspondence: zhixiyang@hit.edu.cn; Tel.: +86-0451-86414883

† These authors contributed equally to this work.

Abstract: The vehicle detection in remote sensing images is a challenging task due to the small size of the objects and interference of a complex background. Traditional methods require a large number of anchor boxes, and the intersection rate between these anchor boxes and an object's real position boxes needs to be high enough. Moreover, the size and aspect ratio of each anchor box need to be designed manually. For small objects, more anchor boxes need to be set. To solve these problems, we regard the small object as a keypoint in the relevant background and propose an anchor-free vehicle detection network (AVD-kpNet) to robustly detect small-sized vehicles in remote sensing images. The AVD-kpNet framework fuses features across layers with a deep layer aggregation architecture, preserving the fine features of small objects. First, considering the correlation between the object and the surrounding background, a 2D Gaussian distribution strategy is adopted to describe the ground truth, instead of a hard label approach. Moreover, we redesign the corresponding focus loss function. Experimental results demonstrate that our method has a higher accuracy for the small-sized vehicle detection task in remote sensing images compared with several advanced methods.

Keywords: vehicle detection; keypoint detection; Gaussian label; deep learning; remote sensing images

check for updates

Citation: Yu, L.; Zhi, X.; Hu, J.; Jiang, S.; Zhang, W.; Chen, W. Small-Sized Vehicle Detection in Remote Sensing Image Based on Keypoint Detection. *Remote Sens.* **2021**, *13*, 4442. https://doi.org/10.3390/rs13214442

Academic Editor: Sander Oude Elberink

Received: 13 October 2021
Accepted: 1 November 2021
Published: 4 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The development of vehicle detection technology in remote sensing images makes it possible to obtain traffic information in time, which is significant for road traffic monitoring, management and scheduling applications. However, in remote sensing images with low resolution, the vehicle object has only a few pixels, with a lack of shape, texture, contour and other features. In addition, the background often interferes with the object information in the detection process. The detection of small-sized vehicle in remote sensing images is a difficult problem.

There are many definitions of small objects; we use the definition by the International Society for Optical Engineering (SPIE), that is, a number of pixels $\leq$80. The current literature for vehicle detection in remote sensing images can be divided into descriptor-based and feature-learning-based methods. The traditional feature-descriptor-based approaches generally consist of three stages: vehicle localization, feature extraction and classification.

Currently, most of the object detection methods are based on the idea of feature detection. Feature generally refers to the part of interest in the image, such as a point, a line segment or a region. Here, feature detection consists in calculating all the positions of the image, converting the image into feature values, and judging whether the feature at a position belongs to a given type. Generally speaking, the proportion of objects in remote sensing images is small, and the background information of remote sensing images is more complex than that of natural images. Object features are affected by background features, which makes it more difficult to detect specific features in remote sensing images, and the

object detection in remote sensing images also faces greater challenges. Therefore, it is important to design robust vehicle detection algorithms suitable for remote scenes.

According to whether the anchor needs to be set, the current object detection methods based on convolutional neural network (CNN) can be divided into an anchor-based method or an anchor-free method. Anchor-based methods need to cover the object well, by setting the bounding box in advance, which makes it possible for a few pixels to produce large errors in small-sized objects annotation. It is especially difficult to judge whether the pixels near the bounding box belong to the object.

Aimed at solving these problems, in this paper, we regard the small object as a keypoint in the relevant background and propose an anchor-free vehicle detection network (AVD-kpNet). The AVD-kpNet framework has a feature extraction part using deep layer aggregation (DLA) [1] and a detection head. Then, we use a 2D Gaussian distribution to describe the ground truth, instead of a hard label, and the corresponding focus loss function is designed for a soft label.

The rest of this paper is organized as follows. Section 2 reviews related work on object detection and applications of CNN in remote sensing images. Section 3 presents the details of the proposed AVD-kpNet framework. Section 4 reports the object detection experiments for the proposed AVD-kpNet framework and Section 5 analyzes the detection performance of the proposed AVD-kpNet framework. Finally, Section 4 concludes this paper.

2. Related Work

A sliding-window approach [2–6] is one of the most widely used methods for vehicle localization. However, many preset parameters such as window size and sliding step size have a great influence on the detection performance. Some studies have also looked at alleviating some of these shortcomings using techniques such as simple linear iterative clustering [7] or edge-weighted centroidal Voronoi tessellations based algorithm [8]. These methods use hard labels, but for the small-sized object detection task in remote sensing images, the relationship between the object and the background is ignored. Moreover, for the small-sized object detection task in remote sensing images, the number of sliding windows is very large, resulting in a lot of redundant computation.

Liu et al. [2] used a sliding window to collect the area to be processed and apply a fast binary detector using integral channel features in a soft-cascade structure. A multiclass classifier was applied to the output of a binary detector to give the direction and type of vehicle. ElMikaty et al. [3] used a sliding-window approach which consisted of a window evaluation, the extraction and encoding of features, classification and postprocessing. Xu et al. [4] improved the vehicle detection performance of the original Viola–Jones detector by rotating the image. Wu et al. [5] proposed an optical remote sensing imagery detector with the processing steps of channel features extraction, feature learning, fast image pyramid matching and boosting strategy. However, these techniques are more suitable for images with large-sized objects in high-resolution remote sensing images (RSIs). For small-sized objects, these methods have difficulty extracting detailed features.

With the rapid development of deep convolution networks, CNNs have shown a good ability for image abstraction. In the field of computer vision, CNN technology has become a research hotspot for semantic segmentation [9–11], object detection [12–16], image classification [17,18] and human pose estimation [19]. CNNs have achieved good results not only in high-level semantic feature detection, but also in low-level feature detection. For example, some researchers used CNNs for edge detection [20–23]. The results showed that CNNs are better than traditional methods for edge detection in natural images. The application of a CNN's ability for image abstraction in detection also greatly promotes the development of object detection.

CNNs can learn features from remote sensing images and detect objects. The literature [24–28] primarily used fast/faster region-based CNN (R-CNN) frameworks to detect vehicles. Chen Huai et al. [29] used transfer-compression learning to train a shallow network, and a location method based on a threshold. Audebert et al. [30] used a segment-

before-detect method for segmentation and subsequent detection and classification of several varieties of vehicles in high-resolution remote sensing images. Koga et al. [31] tailored the correlation alignment domain adaptation (DA) and adversarial DA for the vehicle detector to improve the performance in different data. Mandal et al. [32] constructed a one-stage vehicle detection network by introducing blocks composed of two convolutional layers and five residual feature blocks at multiple scales. The detection method based on region classification has disadvantages. For small-sized objects, more anchors need to be designed manually. Hyper-parameters associated with anchors are usually related to data sets, which are difficult to tune and to be generalized. Taking CornerNet and CenterNet as representatives, the idea of region classification regression is abandoned, and the object position is determined by detecting the center or corner of the object [33,34]. It is not necessary to set anchors when small-sized objects are viewed as keypoints in the related feature map.

Heatmaps have been used in human pose estimation, such as marking each joint position of the human body [19] or different parts [35]. In remote sensing images, there is a certain correlation between the object and the surrounding background. In this paper, we use 2D Gaussian heatmaps to reflect the uncertainty of pixel labels around the center of the object. The ground truth of the heatmap covers the object and the pixels around the object.

3. Proposed Framework

We use a nonstandard two-dimensional Gaussian function to construct the ground truth. The AVD-kpNet framework consists of two parts: a feature extraction network and a detection head. The feature extraction network integrates shallow features with deep features. The detection head predicts the category of each pixel in the output heatmap and the position offset of keypoints. We design a variant of focal loss to penalize the difference between the category of each pixel output by the network and the ground truth. We use an L1 loss function to penalize the keypoints position offset caused by pooling in the network. The output of AVD-kpNet is the heatmap of the object category probability. We determine the location of the object instance according to the maximum value of the neighborhood on the heatmap.

3.1. Overall Architecture

We designed an anchor-free object detector, named AVD-kpNet, to simultaneously perform small-sized vehicle object localization and classification. The proposed detector utilizes the AVD-kpNet framework to learn the salient feature maps from the input image. The entire architecture of the proposed remote sensing object detector is shown in Figure 1.

Due to the small number of pixels of the object, most of the information will be lost if the image is downsampled through the pooling layer many times. Therefore, it is necessary to fuse the information of different resolutions. The backbone network of AVD-kpNet consists of a DLA. DLA combines the advantages of dense connections and feature pyramids, which can aggregate semantic and spatial information. As proven by [36,37] a DLA structure has good performance in the field of detection and recognition.

In order to resolve the feature map to the detection result, a detection head is added after the feature extraction part. The detection head contains a 3×3 and 1×1 convolutional layer, followed by an offset prediction and a heatmap of category predictions. We take the local maximum position as predicted coordinates of objects, as shown in Figure 2.

3.2. Keypoint-Based Prediction Module

Most object detection systems tend to train a deep neural network to extract the deep features of candidate regions, and then predict the class probability of these regions. Using an anchor-based method, it is necessary to design the anchor in advance. For small objects, due to the small number of object pixels, only a very accurate part of the anchor is used as a positive sample. Among a large number of preset anchors, the anchor with a positive sample only accounts for a small part of all the anchors. Some related studies

have shown that the classification score and positioning accuracy of the prediction results of the anchor-based method are not consistent. This inconsistency means that the anchor with inaccurate positioning is selected in the NMS process or when the detection result is selected according to the classification confidence [38]. A small amount of pixel deviation will lead to a decline of the small-sized object positioning accuracy.

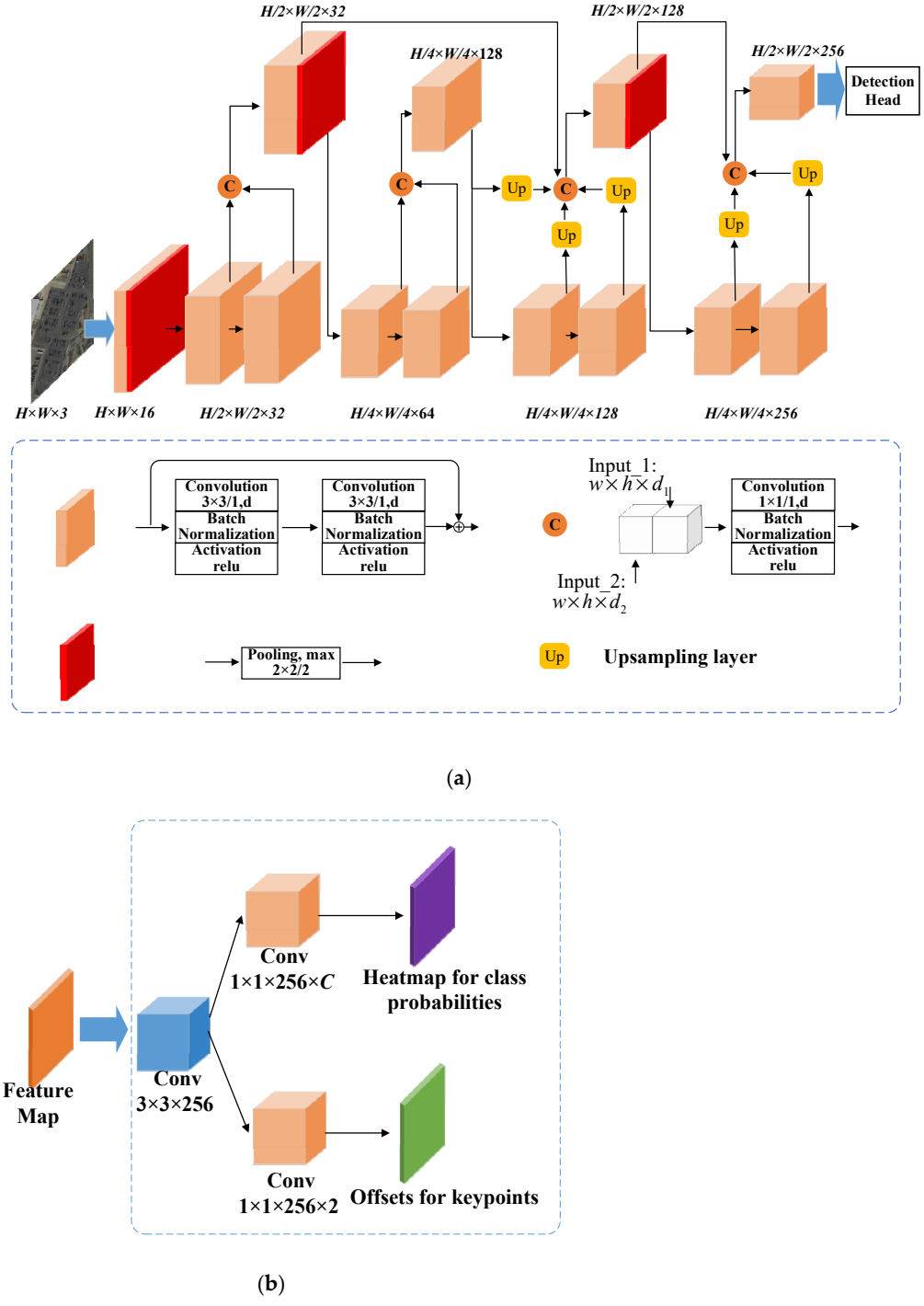


Figure 1. Overall architecture of the AVD-kpNet framework. *H*, *W*: height and width of input image, respectively. *C*: number of object categories. AVD-kpNet mainly comprises the feature extraction part and detection head. The upsampling layer is implemented by fractionally strided convolution. (**a**) Feature extraction part. (**b**) Detection head.

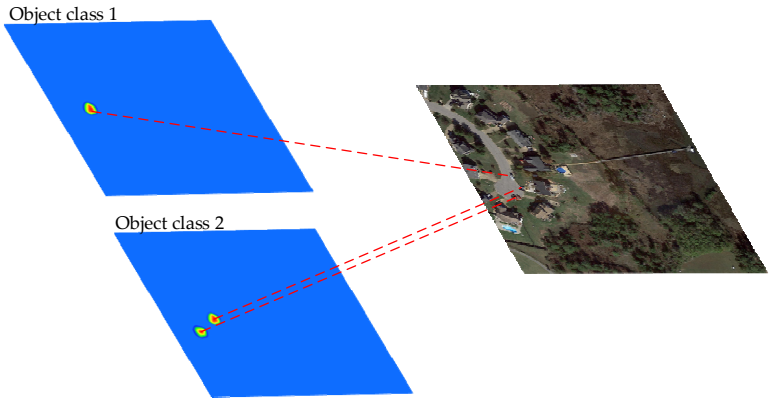


Figure 2. From heatmap to coordinates. The local maximum position is used as the prediction coordinates of the object.

In small-sized object detection, each pixel belonging to the small-sized object has a great influence on the final detection result. In contrast, the conventional scale object has many pixels, even if the loss of a few pixels does not have a great impact on the semantic information of the object. At the same time, since the pixels of small-sized object boundary are fused with the background, the anchor-based method will introduce large errors.

In remote sensing images, vehicle objects usually appear in a specific background; the object is correlated with the surrounding area. Thus, we adopt a 2D Gaussian heatmap to label the center of the object as the keypoint, and the closer to the center, the greater the confidence probability of the object. In the heatmap of the same keypoint, some pixels belonging to the object and background are marked at the same time. In our proposal, heatmaps are set to be ground truths at the stage of network training. Each heatmap contains the ground truth of all instances of the same category with the normalized Gaussian function whose parameters vary with the size of objects. The ground truth of the object is given by a nonstandard two-dimensional Gaussian function.

$$Y_{xyc} = \exp\left(-\frac{(x-p_x)^2 + (y-p_y)^2}{2\sigma_p^2}\right) \tag{1}$$

where $(x - p_x)$ and $(y - p_y)$ represent the distances to the center of the object, σ_p is the radius of Gaussian kernel and the calculation formula is:

$$\sigma_p = avg(w, h) \tag{2}$$

where $avg(w, h)$ is the average width and height of the minimum bounding rectangle of the object. This makes the ground truth value around the center of the object is not zero. The ground truth describes the distance relationship between other pixels and the keypoint.

The size of the predicted heatmap is $W \times H \times C$, and C represents the number of categories. If two Gaussian functions overlap for the same class C, we choose the one with the largest element level.

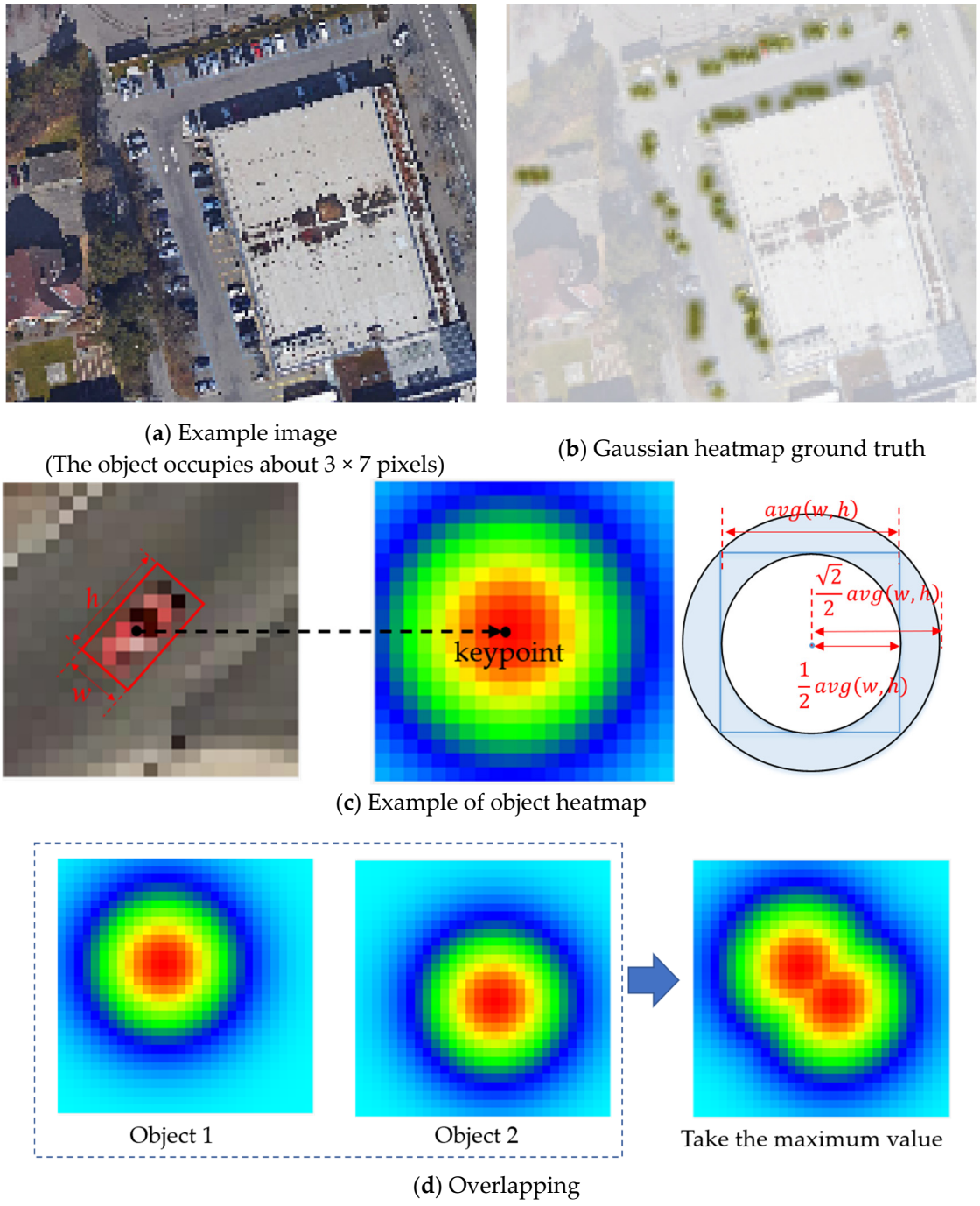
As can be seen from Figure 3, our annotation method regards the target and the background around the target as positive samples of different degrees at the same time.

The heatmaps of objects with the same $avg(w, h)$ are identical. We consider a square region with the length of $avg(w, h)$. When the predicted position and the ground truth of the keypoint satisfy Equation (3), the Gaussian function value of the ground truth is 0.78 and we define the position as close to the object keypoint. When the predicted position and the ground truth of the keypoint satisfy Equation (4), the Gaussian function value of

the ground truth is 0.61 and we define the position as far from the object keypoint. The threshold range can be used to judge whether the detection is correct in the experiment.

$$\begin{cases} (x - p_x) \leq \frac{1}{2} avg(w, h) \\ (y - p_y) \leq \frac{1}{2} avg(w, h) \end{cases} \tag{3}$$

$$\begin{cases} (x - p_x) \geq \frac{\sqrt{2}}{2} avg(w, h) \\ (y - p_y) \geq \frac{\sqrt{2}}{2} avg(w, h) \end{cases} \tag{4}$$



(**a**) Example image (The object occupies about 3 × 7 pixels)

(**b**) Gaussian heatmap ground truth

(**c**) Example of object heatmap

(**d**) Overlapping

Figure 3. Gaussian heatmap. (**a**) An example image with some small-sized vehicles. (**b**) Gaussian heatmap ground truth of the image (**a**). (**c**) An example of object heatmap, and the threshold of the distance from the keypoint. (**d**) When the Gaussian functions belonging to the same class of objects overlap, the maximum value is taken.

3.3. Loss Function

Let p_{cij} be the score at location (i, j) for class c in the predicted heatmap and let g_{cij} be the score at location (i, j) for class c in the ground truth heatmap. We design a variant of focal loss:

$$L_{point} = -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|g_{cij} - p_{cij}\right|^{\alpha}\Big(\left(1 - g_{cij}\right)^{\beta}\log\left(1 - p_{cij}\right) + {g_{cij}}^{\beta}\log\left(p_{cij}\right)\Big); g_{cij} \in [0, 1] \quad (5)$$

where N is the number of objects in an image, α is used to limit the dominance of gradients caused by easy examples, β is the hyper-parameter that controls the contribution of each keypoint. For setting the parameter values, we refer to the [39], and perform fine-tuning during the experiment to obtain the final parameters. We set α to 4 and β to 2 in all experiments.

When g_{cij} is 0 or 1, the formula for L_{point} degenerates to a general focus loss:

$$L_{point}\begin{cases} -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|1 - p_{cij}\right|^{\alpha}\log\left(p_{cij}\right) & g_{cij} = 1 \\ -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|p_{cij}\right|^{\alpha}\log\left(1 - p_{cij}\right) & otherwise \end{cases} \quad (6)$$

After image downsampling, the keypoint position of the ground truth will produce a certain deviation, and we predict a local offset for each center point. This offset is trained by the L1 loss function.

$$L_{off} = \frac{1}{N}\sum_{p}\left|O_p - \left(\frac{g}{d_r} - p\right)\right| \quad (7)$$

where O_p is the prediction of the offset, d_r is the downsampling rate of the network, g is the center position of the real heatmap of the object and also the maximum position of the heatmap value. p is the local maximum position of the predicted heatmap.

In conclusion, the total loss function is:

$$L = Loff_{offpoint} \quad (8)$$

where λ_{off} is the weight of L_{offset}. Similar to the [33], we set λ_{off} to 1.

4. Experiments Results and Discussion

4.1. Data Set

We collected 877 high-resolution remote sensing images from DOTA [40] and Google Earth. The original size range of the images vary from 800 × 800 pixels to 4000 × 4000 pixels. The images include roads, trees, houses and other kinds of backgrounds. The spatial resolution ranges from 0.1 m to 0.3 m. We resampled these high-resolution remote sensing images with a downscaling factor of 5, so that the number of vehicle object pixels is less than 80, which conforms to the definition of small-sized objects in this paper. We cropped all the images to 512 × 512 pixels, then, we marked the ground truth of the image by the heatmap method illustrated in Section 3.2. We divided vehicle objects into two categories, one of which mainly includes sedans, hatchbacks and similar sized cars with an actual length of around 4 meters, and we called it vehicle_1. The other category mainly included container cars, buses, and the actual length is more than 10 m, which we called vehicle_2. There were about 1700 instances of vehicle_1 and 1500 instances of vehicle_2. The average pixel size of vehicle_1 and vehicle_2 in each resampled image was 2 × 4~5 × 10 and 2 × 9~4 × 20, respectively. Figure 4 shows some examples of the employed data set.

In our experiments, we divided these images into three sets, 1033 images as the training set, 295 images as the validation set and 149 images as the test set. The resolution of the images was 512 × 512 pixels. For the training of AVD-kpNet, we used the commonly used data normalization method, zero-mean normalization, which subtracts the average

value of training data from the input image and then divides it by the standard deviation. Moreover, we used mosaic data augmentation. We cropped the images into small images with a size of 128 ×128 pixels; the sliding-step was 128 with a random shift from 0 to 32 pixels. After stitching 16 pictures, we obtained a new picture, and then we fed that picture to the neural network for training. This preprocessing makes the input image contain a richer background. In addition, this significantly reduces the need for a large mini-batch size. This data enhancement method has been applied in YOLOv4 [41] and achieved good results. For the inference, the input image was not cropped. Several augmentation technologies were also adopted to increase the amount of data. We randomly flipped images horizontally and vertically, and an angle was randomly chosen from (90°, 180°, 270°) to rotate the images. The normalization and augmentation sequences from the training process are shown in Figure 5.



Figure 4. Examples of images from the data set. Images contain vehicle_1 and vehicle_2 objects, as well as roads, trees, houses and other kinds of backgrounds.

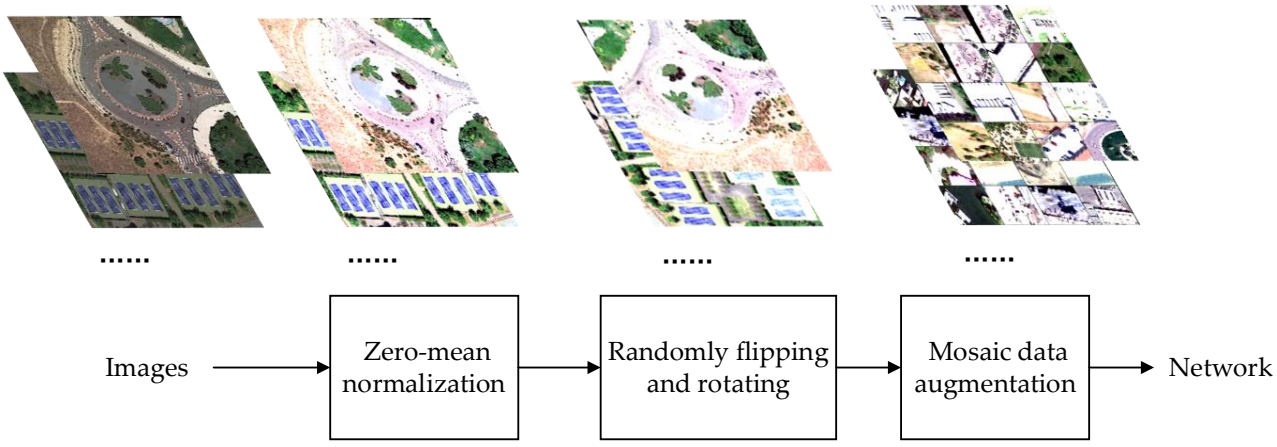


Figure 5. Data normalization and augmentation during training.

4.2. Evaluation Metrics

In the inference stage, the local maximum value of the heatmap is taken as the keypoint position. As can be seen from Figure 6, all the response points on the heatmap are compared with their eight adjacent points. If the response value of the point is greater than or equal to its eight adjacent points, it is retained. After finding the maximum position of the neighborhood, only the loss of offset is calculated for this point.

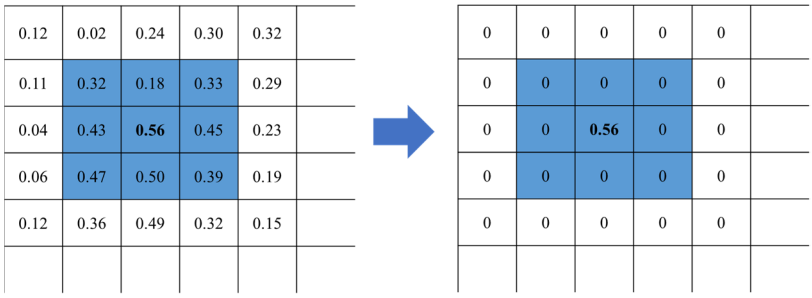


Figure 6. The local maximum value is retained, and other positions are assigned to 0.

The real distance (we set the value as D) between the predicted coordinate position and the ground truth is used to judge the consistency between the predicted position and the actual situation. Then, we set a threshold to judge whether the position prediction is accurate. In our experiments, the category prediction is considered as true positive (TP) if keypoint similarity (OKS) is above 0.78 while it is considered to be false positive (FP) if OKS is less than 0.78. If an object is not detected and recognized, it is considered a false negative (FN).

$$OKS = e^{-\frac{D^2}{2(avg(w,h))^2}} \tag{9}$$

where $avg(w,h)$ is the average width and height of the minimum bounding rectangle of the object.

We use average precision (AP) and mean average precision (mAP) to measure the performance of the detection model.

$$precision = \frac{TP}{TP + FP} \tag{10}$$

$$recall = \frac{TP}{TP + FN} \tag{11}$$

$$AP = \int_0^1 precision(recall)d(recall) \tag{12}$$

$$mAP = \frac{AP_{Vehicle_1} + AP_{Vehicle_2}}{2} \tag{13}$$

The false alarm rate (FAR) is an important evaluation metric in practical tasks. If the FAR is too high, it will have a negative impact on practical application tasks. Therefore, the FAR is introduced to further evaluate the performance of the detection model:

$$FAR = \frac{\text{number of detected false alarms}}{\text{number of detected candidates}} \tag{14}$$

4.3. Experiment Results and Analysis

For setting the parameters of the training process, we refer to [1], and fine-tuned them in the experiment. We set the learning rate to 0.01, the epoch to 150, the batch size to 8, the stochastic gradient descent (SGD) with a momentum of 0.9 and the weight decay to 10^{-4}. We started the training with a learning rate of 0.01, reducing 10 times every 50 epochs. We

did not use pretrained weights. Figure 7 shows some examples of detection results, and Figure 8 shows some examples of enlarged images of some smaller objects.

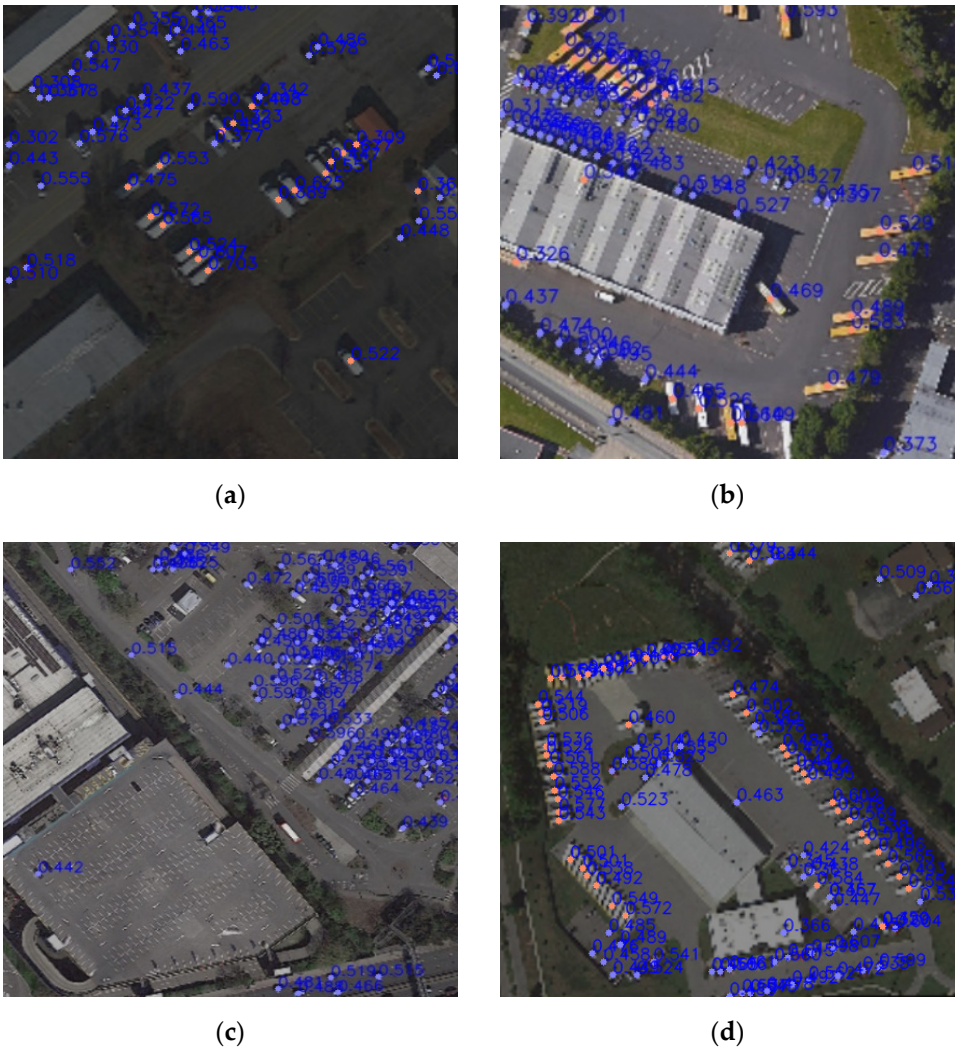
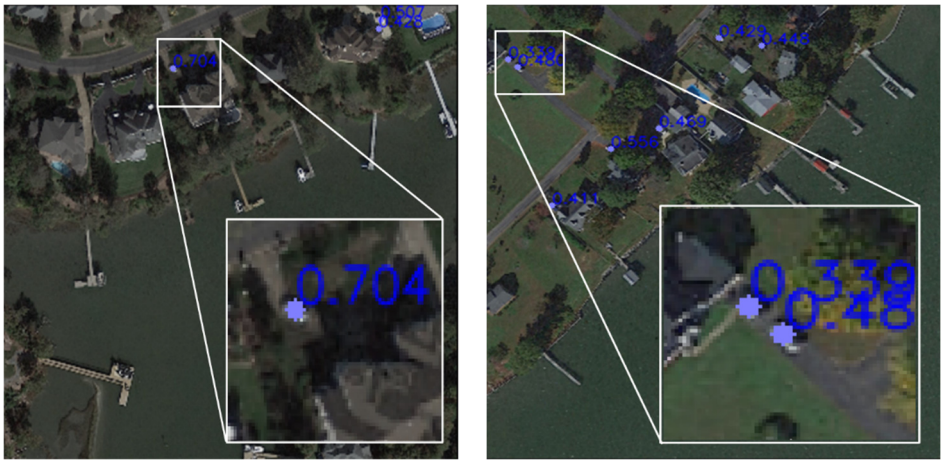


Figure 7. Examples of detection results. The two kinds of objects are represented by blue and orange dots respectively. (**a**) Example 1, (**b**) Example 2 (**c**) Example 3, (**d**) Example 4.

The test results of other networks were calculated under the condition of Intersection over Union (IoU) $\geq$0.5. Table 1 reports the evaluation values of *AP*, *mAP* and *FAR* for different methods.

Table 1. Comparisons of different methods on two kinds of object. The bold value indicates the best performance of each term.

Method	Vehicle_1 (AP)	Vehicle_2 (AP)	mAP	FAR
Faster R-CNN	10.0%	13.4%	11.70%	72.1%
YOLOv3_416 × 416	17.6%	38.3%	27.95%	65.0%
YOLOv4	21.9%	42.3%	32.10%	49.5%
RetinaNet	17.0%	39.2%	28.10%	60.3%
AVD-kpNet (hard label + original focal loss)	33.4%	38.2%	35.80%	45.2%
AVD-kpNet	**38.2%**	**49.1%**	**43.65%**	**38.3%**



(**a**) (**b**)

Figure 8. Examples of enlarged images of some smaller objects. The object occupies about 3 × 7 pixels. (**a**) Example 1, (**b**) Example 2.

In order to verify the contribution of the proposed soft label and the new loss function, ablation experiments were carried out on the existing data sets. The means of hard labels in Table 1 consisted of marking with a circle whose radius was the average length and width of the object. We considered a positive sample when inside the circle, set by 1, and a negative sample when outside the circle, set by 0. Without changing the network structure, using hard labels and original focal loss, the *mAP* value is 35.80% and *FAR* value is 45.2%. Compared with the YOLOv4 model based on anchor boxes, the performance of the detection model is slightly improved, the *mAP* value is increased by 3.70%, and the *FAR* value is reduced by 4.3%. When the detection model uses soft labels and the proposed loss function, the performance of the detection model is further improved; the *mAP* value is increased by 7.85% and the *FAR* value is reduced by 6.9%.

In order to analyze the impact of relatively small training data sets on our algorithm, we used 2/3, 1/2, 1/4 and 1/8 training data sets for model training, respectively. For models trained with training data sets of different sizes, complete test data sets were used for model testing, and the test results are shown in Table 2. It can be seen from the Table 2 that the detection performance of the model gradually improves with the increase of the training data set. The detection model using 2/3 of the training data can also approach the performance when using all the training data.

Table 2. Comparisons of different data set scale. The bold value indicates the best performance of each term.

Training Data Set Scale	Vehicle_1 (AP)	Vehicle_2 (AP)	mAP	FAR
1/8	11.5%	12.4%	11.95%	54.7%
1/4	24.6%	32.9%	28.75%	50.8%
1/2	35.6%	44.6%	40.10%	42.4%
2/3	38.1%	48.8%	43.45%	38.6%
All	**38.2%**	**49.1%**	**43.65%**	**38.3%**

Based on the comprehensive analysis results, we can obtain the following conclusions: (1) in the experimental indexes, AVD-kpNet obtains the best AP, mAP and FAR; (2) the detection performance for vehicle_2 is better than for vehicle_1, possibly because the details of vehicle_1 instances are not good enough to be detected; (3) for the same network structure, the model with a Gaussian heatmap as ground truth shows better performance. For the AVD-kpNet framework with hard labels, if the radius of the circle is too small, it cannot completely cover the object, resulting in the uncovered part being used as the

background. If the radius of the circle is too large, many background pixels will be used as the object.

5. Conclusions

In this paper, a framework named AVD-kpNet was proposed for small-sized vehicle detection in remote sensing images. We regarded small-sized objects as keypoints in the background. The outputs of the AVD-kpNet framework are heatmap layers of objects and we can get the positions of object instances by locating peak regions in the heatmap layers. We proposed a method to get the ground truth of vehicles in remote sensing image using a 2D Gaussian distribution. The purpose of the marking method was to construct the relationship between small-sized objects and the surrounding background, so as to improve the detection ability of keypoints. Moreover, we designed a variant of focal loss to reduce the impact of easy examples. Comparing our proposal with several other methods, we draw the conclusion that the method based on AVD-kpNet can achieve state-of-the-art results. In further research, we will improve the robustness of the model for the detection problem when the object is partially occluded.

Author Contributions: Conceptualization, L.Y.; methodology, L.Y.; software, L.Y.; validation, X.Z.; formal analysis, L.Y.; investigation, J.H.; resources, J.H. and W.C.; data curation, X.Z.; writing—original draft preparation, L.Y.; writing—review and editing, J.H.; visualization, S.J. and W.C.; supervision, S.J.; project administration, W.Z.; funding acquisition, W.Z. 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, grant number: 61975043.

Data Availability Statement: The dataset for this research is released in November 2017, which is available at https://captain-whu.github.io/DOTA/dataset.html, accessed on 30 October 2021.

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

References

1. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep layer aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2403–2412.
2. Liu, K.; Mattyus, G. Fast Multiclass Vehicle Detection on Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1938–1942. [CrossRef]
3. ElMikaty, M.; Stathaki, T. Detection of Cars in High-Resolution Aerial Images of Complex Urban Environments. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5913–5924. [CrossRef]
4. Xu, Y.; Yu, G.; Wu, X.; Wang, Y.; Ma, Y. An Enhanced Viola-Jones Vehicle Detection Method From Unmanned Aerial Vehicles Imagery. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1845–1856. [CrossRef]
5. Zhou, H.; Wei, L.; Lim, C.P.; Creighton, D.; Nahavandi, S. Robust Vehicle Detection in Aerial Images Using Bag-of-Words and Orientation Aware Scanning. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7074–7085. [CrossRef]
6. Wu, X.; Hong, D.; Tian, J.; Chanussot, J.; Li, W.; Tao, R. ORSIm Detector: A Novel Object Detection Framework in Optical Remote Sensing Imagery Using Spatial-Frequency Channel Features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5146–5158. [CrossRef]
7. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Suesstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2281. [CrossRef] [PubMed]
8. Wang, J.; Wang, X. VCells: Simple and Efficient Superpixels Using Edge-Weighted Centroidal Voronoi Tessellations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1241–1247. [CrossRef] [PubMed]
9. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1520–1528.
10. Dai, J.; He, K.; Sun, J. Instance-aware Semantic Segmentation via Multi-task Network Cascades. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3150–3158.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision-Eccv 2016, Pt I*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9905, pp. 21–37.
13. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.

14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
16. Hu, J.; Zhi, X.; Shi, T.; Zhang, W.; Cui, Y.; Zhao, S.J.R.S. PAG-YOLO: A Portable Attention-Guided YOLO Network for Small Ship Detection. *Remote Sens.* **2021**, *13*, 3059. [CrossRef]
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
18. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 30th Ieee Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 2261–2269.
19. Pfister, T.; Charles, J.; Zisserman, A. Flowing ConvNets for Human Pose Estimation in Videos. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1913–1921.
20. Bertasius, G.; Shi, J.; Torresani, L. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4380–4389.
21. Liu, Y.; Cheng, M.-M.; Hu, X.; Wang, K.; Bai, X. Richer convolutional features for edge detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3000–3009.
22. Shen, W.; Wang, X.; Wang, Y.; Bai, X.; Zhang, Z. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3982–3991.
23. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1395–1403.
24. Sakla, W.; Konjevod, G.; Mundhenk, T.N. Deep Multi-Modal Vehicle Detection in Aerial ISR Imagery. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision, Santa Rosa, CA, USA, 24–31 March 2017; pp. 916–923.
25. 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, Santa Rosa, CA, USA, 24–31 March 2017; pp. 311–319.
26. Nie, K.; Sommer, L.; Schumann, A.; Beyerer, J. Semantic Labeling based Vehicle Detection in Aerial Imagery. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 626–634.
27. 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]
28. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
29. Chen, H.; Zhang, L.; Ma, J.; Zhang, J.J.N. Target heat-map network: An end-to-end deep network for target detection in remote sensing images. *Neurocomputing* **2019**, *331*, 375–387. [CrossRef]
30. Audebert, N.; Le Saux, B.; Lefèvre, S.J.R.S. Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sens.* **2017**, *9*, 368. [CrossRef]
31. Koga, Y.; Miyazaki, H.; Shibasaki, R. A Method for Vehicle Detection in High-Resolution Satellite Images that Uses a Region-Based Object Detector and Unsupervised Domain Adaptation. *Remote Sens.* **2020**, *12*, 575. [CrossRef]
32. Mandal, M.; Shah, M.; Meena, P.; Devi, S.; Vipparthi, S.K.J.I.G.; Letters, R.S. AVDNet: A small-sized vehicle detection network for aerial visual data. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 494–498. [CrossRef]
33. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6569–6578.
34. 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.
35. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1840.
36. Cui, Y.; Song, Y.; Sun, C.; Howard, A.; Belongie, S. Large scale fine-grained categorization and domain-specific transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4109–4118.
37. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 113–123.
38. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L.J.A.P.A. Dynamic anchor learning for arbitrary-oriented object detection. *arXiv* **2020**, arXiv:2012.041501, 6.
39. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.

40. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
41. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M.J.A.P.A. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.