remote sensing

MDPI

Article

RelationRS: Relationship Representation Network for Object Detection in Aerial Images

Zhiming Liu [1,2], Xuefei Zhang [1,2,*], Chongyang Liu [1], Hao Wang [3], Chao Sun [1], Bin Li [1], Pu Huang [1,2], Qingjun Li [1], Yu Liu [1], Haipeng Kuang [1,2] and Jihong Xiu [1,2]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; liuzhiming@ciomp.ac.cn (Z.L.); liuchongyang@ciomp.ac.cn (C.L.); sunchao@ciomp.ac.cn (C.S.); libin@ciomp.ac.cn (B.L.); huangpu@ciomp.ac.cn (P.H.); liqingjun@ciomp.ac.cn (Q.L.); liuyu@ciomp.ac.cn (Y.L.); kuanghaipeng@ciomp.ac.cn (H.K.); xiujihong@ciomp.ac.cn (J.X.)

[2] Key Laboratory of Airborne Optical Imaging and Measurement, Chinese Academy of Sciences, Changchun 130033, China

[3] Electronic Information Engineering College, Changchun University, Changchun 130022, China; wanghao@ciomp.ac.cn

* Correspondence: zhangxuefei@ciomp.ac.cn; Tel.: +86-135-0442-2784

Abstract: Object detection is a basic and important task in the field of aerial image processing and has gained much attention in computer vision. However, previous aerial image object-detection approaches have insufficient use of scene semantic information between different regions of large-scale aerial images. In addition, complex background and scale changes make it difficult to improve detection accuracy. To address these issues, we propose a relationship representation network for object detection in aerial images (RelationRS): (1) Firstly, multi-scale features are fused and enhanced by a dual relationship module (DRM) with conditional convolution. The dual relationship module learns the potential relationship between features of different scales and learns the relationship between different scenes from different patches in a same iteration. In addition, the dual relationship module dynamically generates parameters to guide the fusion of multi-scale features. (2) Secondly, the bridging visual representations module (BVR) is introduced into the field of aerial images to improve the object detection effect in images with complex backgrounds. Experiments with a publicly available object detection dataset for aerial images demonstrate that the proposed RelationRS achieves a state-of-the-art detection performance.

Keywords: object detection; aerial imagery; conditional convolution; relationship representation; bridging visual representations

Citation: Liu, Z.; Zhang, X.; Liu, C.; Wang, H.; Sun, C.; Li, B.; Huang, P.; Li, Q.; Liu, Y.; Kuang, H.; Xiu, J. RelationRS: Relationship Representation Network for Object Detection in Aerial Images. *Remote Sens.* **2022**, *14*, 1862. https://doi.org/10.3390/rs14081862

Received: 14 February 2022
Accepted: 4 April 2022
Published: 13 April 2022

1. Introduction

The technique of object detection in aerial images refers to extracting the positions of the objects defined by domain experts, according to whether a kind of object is common and its value for real-world applications [1,2]. With the rapid development of remote sensing acquisition technology, object detection methods in aerial images are widely used in ship monitoring, maritime search and rescue, traffic control, power line inspection, military reconnaissance, and other fields [3]. In addition, for unmanned aerial vehicles (UAVs) and satellites with limited energy resources, aerial image object detection technology is also used as a pre-processing method to select important images with suspected targets and return them to the ground station first. This puts higher requirements on the accuracy and speed of object-detection algorithms [4]. Based on this rapidly growing demand, traditional object-detection algorithms using the logic of extracting hand-designed low-level features, and then classifying and predicting bounding boxes can no longer meet the detection tasks under fast, multiple categories and complex background conditions [5–7].

Widely used convolutional neural networks (CNNs) have greatly promoted the development of object-detection technology. Overall, the current object detection networks can be divided into single-stage detectors and two-stage detectors. On the one hand, the two-stage detector method can be also called the region proposal-based method. This type of algorithm usually uses a small two-class network to extract the region proposals firstly, and then performs a bounding boxes tuning procedure and category prediction. The region proposal-based method can achieve higher detection accuracy, but its structure is not suitable for edge computing devices. Deformable part model (DPM) [8], region-based convolutional neural network (R-CNN) [9], Fast R-CNN [10], Faster R-CNN [11], feature pyramid network (FPN) [12], Cascade R-CNN [13], Mask R-CNN [14], etc., are the representative algorithms. On the other hand, the single-stage detector can be also called the regression-based method, including Overfeat [15], you only look once (YOLO) [16], YOLOv2 [17], YOLOv3 [18], YOLOv4 [19], YOLOX [20], single shot detector (SSD) [21], fully convolutional one-stage object detection (FCOS) [22], and RetinaNet [23]. The single-stage detector performs object detection on an entire image at one time, and predicts the bounding boxes as the coordinate regression task. This kind of method is very suitable for the embedding of edge computing devices, but its accuracy is not as good as the two-stage detector.

It is worth mentioning that some methods transform the prediction of the bounding boxes into a task of predicting the key points of the bounding boxes [24–26]. These methods are fast and efficient, but have difficulty in scenes where the objects obscure each other in natural images.

At present, encouraged by the great success of deep-learning-based object detection in natural images [27,28], many researchers have proposed utilizing a similar methodology for object detection in aerial images [7,29]. Common natural images usually come from surveillance cameras, mobile phones, cameras, and other imaging platforms, and the imaging procedure is often carried out from a head-up perspective or a head-up perspective with a small angle. Scenes in natural images are usually relatively single. This means that there is no strict logical relationship between people and vehicles and the background. Vehicles can appear in high-rise buildings, and vehicles can also appear on the beach. In addition, sizes of objects strictly follow the specification of near–large and far–small. Remote sensing images usually come from satellites, aircraft, fixed-wing unmanned aerial vehicles (UAVs), rotary-wing UAVs, and other platforms. When imaging, a vertical view perspective or tilt photography mechanism can be adopted. As a single image has a larger size, it covers a larger geographical range and includes more types of ground objects, with a more complex background. This means a single image can often cover a complete scene type, such as an airport. Different scenes and their unique objects constitute scene semantics, such as airplanes and airports, vehicles and roads, cities and viaducts, etc. This is very different from the representation of scenes in natural images. Figure 1 shows the difference between natural images and aerial images. First of all, the backgrounds of aerial images are more complex, which put forward higher requirements on algorithms. Secondly, the aerial images are taken from the vertical view perspective when acquiring, so there is almost no occlusion between the objects. In addition, aerial images have rich scene-target semantic information. Finally, due to the different acquisition platforms, flight trajectories, and sensors used when acquiring aerial images, almost every aerial image product has unique resolution and imaging characteristics. This leads to drastic scale changes of the same object. Moreover, the scales between different objects are also quite different. To sum up, it is impossible to directly apply the common object-detection algorithms based on deep learning from the field of natural images to the field of aerial image object detection.

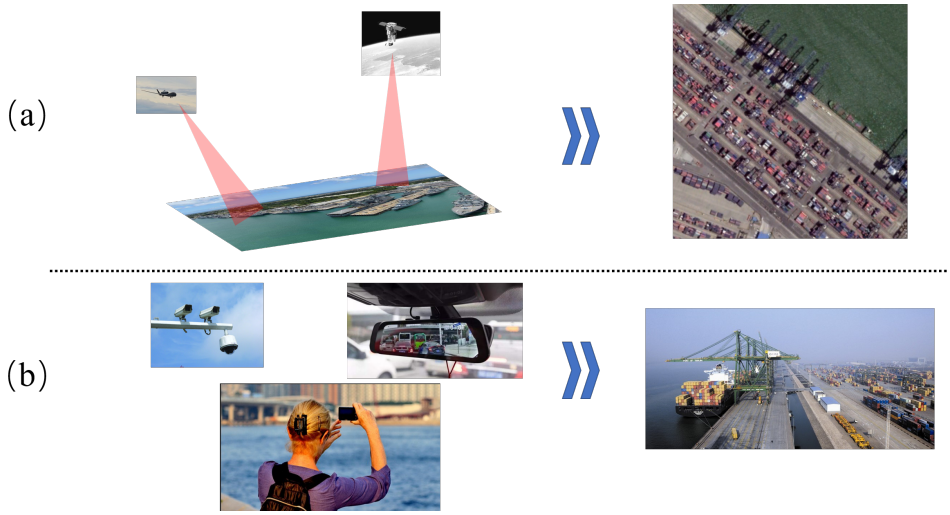

Figure 1. The difference between natural images and aerial images. (**a**) A port image data obtained by aerospace platforms or aviation platforms. (**b**) A port image data captured by surveillance equipment or handheld devices.

In this paper, in order to obtain improved object detection results with optical remote sensing images, we followed these steps: (1) We put forward a fast and effective network structure based on one-stage detector algorithms, namely, relationship representation network for object detection in aerial images (RelationRS). In RelationRS, to solve the problem of dramatic scale changes in aerial images, a novel dual relationship module has been designed to realize the extraction and characterization of multi-scale relations and the potential relationships between different scenes from one iteration. The dual relationship module can guide the multi-scale features fusion with weight parameters dynamically generated by conditional convolutions. Based on FPN [12] algorithm, conditional convolution mechanism has been used to learn the above two kinds of potential relations and dynamically guide the fusion of multi-scale feature maps. The algorithm of dual relationship module is different from some methods [2,30] based on extra semantic branches, and can avoid the serious speed loss caused by extra branches while improving the overall accuracy of the network. (2) In order to cope with the complex background of aerial images, the bridging visual representation module (BVR) [31] used in natural image object-detection task is introduced into the aerial image object-detection task for the first time to fuse the key-point representation algorithm and the bounding box representation algorithm. Based on the imaging characteristics of vertical view perspective, there is almost no occlusion between the objects, so the key-point representation can be introduced into the rectangular bounding box frame. Therefore, the two object representation methods can be combined simultaneously to improve the positioning accuracy of objects and reduce the influence of complex background.

The proposed algorithm tested on the DOTA dataset [1] demonstrates that the proposed RelationRS achieves a state-of-the-art detection performance.

The rest of this paper is organized as follows. Section 2 gives a brief review of the related work on aerial image object detection based on deep learning, the multi-scale feature representations, and the conditional convolution mechanism. In Section 3, we introduce the proposed method in detail. The details of the dataset, the experiments conducted in this study, and the results of the experiments are presented in Section 4. Section 5 concludes this paper with a discussion of the results.

2. Related Work

2.1. Object Detection of Aerial Images

Given the characteristics of aerial images, many researchers currently focus on the two-stage detector as the baseline to obtain higher-precision detection results, compared to the single-stage detector.

Zou et al. [32] proposed a singular value decomposition network (SVDNet) for ship object detection in aerial images. To improve detection accuracy, the SVDNet combines singular value decomposition and convolutional neural networks. Since the region proposals are extracted by the selective search algorithm [33,34], this non-end-to-end implementation greatly reduces the speed of the algorithm. Dong et al. [35] optimized the non-maximum suppression algorithm, and the proposed Sig-non-maximum suppression method has better accuracy for small objects. Both Deng et al. [36] and Xiao et al. [37] have used sliding windows to extract region proposals, then used convolutional neural networks to extract deep-learning-based features, and finally used non-rotating rectangular boxes to characterize objects. A large number of regions in aerial images belong to the background class where no object exists. Therefore, the method of using the sliding windows to extract region proposals is inefficient, and it is easy to cause an imbalance between positive and negative samples. Xu et al. [38] and Ren et al. [39] applied deformable convolution [40] to aerial image object detection tasks with complex boundaries. Through expanding the receptive field of the convolution kernel, the deformable convolution with variable sampling position is beneficial to extract the complex boundaries of objects in aerial images. Similarly, Hamaguchi et al. [41] has used the dilated convolution filters [42] to optimize the receptive field size of the convolution kernel in aerial image object detection and segmentation, which is conducive to the extraction of scene semantic information. Context-driven detection network (CDD-Net) [43] has used an attention mechanism to improve the ability of multi-scale feature representation. In addition, CDD-Net has used a local semantic feature network to extract the feature information of the area around an object, to compensate for the limited filter size of the CNNs. Finally, the detection task is finished with both the feature maps of region proposals and the information around objects. To improve the efficiency of multi-scale feature utilization, adaptively aspect ratio multi-scale network (A2RMNet) [44] adopted a gate structure to automatically select and fuse multi-scale features and then adopted the same region proposal feature map pooling method as rotational region convolutional neural network (R2CNN) [45] to obtain three different aspect ratios. Feature maps and the attention mechanism are also used to optimize and merge the three feature maps to improve the accuracy of object detection.

Different from the algorithms using the rectangular bounding box frame, some aerial image object detection methods based on two-stage detectors believe that the use of the rotating bounding box frame can better describe the object and reduce the influence of background pixels. Although a rotation region proposal network (RRPN) [46] is proposed for the task of text detection, it has achieved excellent results in many competitions related to aerial imagery. Similarly, both Li et al. [47] and Ding et al. [48] adopted the lightweight structure to obtain more accurate rectangular bounding box characterization with different angles in the region proposal extraction stage. When performing RoI Pooling operations, this kind of rotation bounding box frame tries to eliminate the background pixels in the region proposal to improve the characteristics of the object itself. Yang et al. [49] constructed a novel multi-category rotation detector for small, cluttered and rotated objects (SCRDet). Firstly, for SCRDet, a new feature fusion branch (SF-Net) is proposed to improve the recall value of the region proposals. Then, SCRDet adopted a supervised multi-dimensional attention network (MDA-Net) with self-attention mechanism to reduce the interference of background pixels and improve the feature representation ability of the object itself. Finally, the loss function is improved with a constant factor. Both context-aware detection network (CAD-Net) [30] and global-local saliency constraint network (GLS-Net) [2] improved feature representation ability of the object itself by global semantic branches with attention mechanism or saliency algorithm. These global semantic branches

are used to compensate for the lack of scene information caused by the limited size of the receptive field. In fact, saliency algorithm is one of the important ways to combat complex background. Fang et al. [50] proposed a stereoscopic video saliency detection method based on 3D convolutional neural networks, namely, deep 3D video saliency (Deep3DSaliency). Deep3DSaliency makes full use of the spatio–temporal information between frames to make the target regions more prominent and produce significant differences from the background. Similarly, Jian et al. [51] proposed an effective visual saliency-detection model based on spatial position prior of attractive objects and sparse background features.

Li et al. [52] proposed a network with refine feature pyramid network and multi-layer attention network (RADet). RADet improved the feature representation ability for scale changes by fusing the front-layer feature maps and the deeper-layer feature maps; this kind of fusion strategy is useful for the small-object-detection tasks. Moreover, on the basis of on Mask R-CNN [14], RADet can predict the instance mask and the rotating rectangular bounding box at the same time based on the attention mechanism. In RADet, the minimum rectangular area of the object is used as the truth of the mask, which actually still contains the background area. For this reason, RADet does not realize the high-precision characterization of the object itself. Similarly, inspired by Mask R-CNN, mask-oriented bounding box representation (Mask OBB) [53] adopted the inception module [54] to fuse feature maps from different depths and utilized the attention mechanism to construct semantic segmentation feature maps. Finally, Mask OBB can predict the bounding boxes, rotated bounding boxes, and instance masks simultaneously. In order to deal with the problem of scale changes and small objects, Li et al. [55] proposed a method that can predict bounding boxes and rotated bounding boxes with a module similar to the inception module and a semantic segmentation module. However, the above two methods require multiple types of samples in the training stage. Xu et al. [56] believe that it is difficult to directly predict the rotated bounding boxes, and the detection of the rotated objects should to be implemented step by step; that is, the non-rotated bounding boxes need to be predicted first, and then the offset of the four vertices can be calculated. Zhu et al. [57] proposed that it is difficult and inefficient to directly calculate the angles of the bounding boxes. Therefore, two different two-dimensional periodic vectors are used to represent angles, and a new intersection over union (IoU) is used to solve the problem of the object with a large length and width ratio. Fu et al. [58] extracted features of region proposals with angles and merged feature maps from different depths through bidirectional information flow to enhance the representation ability of the feature pyramid network. Rotation-equivariant Detector (ReDet) [59] encodes the rotation-equivariance and the rotation-invariance, which can reduce the demand for network parameters while realizing the detection of rotated objects. Based on the characteristics of images in frequency, octave convolution-based semantic attention feature pyramid network (OcSaFPN) [3] is proposed to improve the accuracy of object detection in aerial images with noise.

Object-detection algorithms based on single-stage detectors in aerial images have developed rapidly, and have narrowed or even surpassed the accuracy gap with two-stage detectors. The aerial image object-detection algorithms based on single-stage detectors have the common characteristics of being fast, concise, and conducive to the deployment of dedicated computing chips and edge computing equipment. For this reason, object-detection algorithms based on single-stage detectors in aerial images have great potential in production and application in a variety of scenarios. By adding a new branch for scene prediction, you Only Look Twice (YOLT) [60] realizes the simultaneous prediction of the objects and scene information, and forms the association between the object and the scene information from the loss function. However, this method does not make up for the shortcomings of YOLOv2 itself, and most of the datasets in aerial image field lack scene labeling information. Inspired by SSD, feature-merged single-shot detection method (FMSSD) [61] adopted the dilated convolutional filter to enlarge the size of the receptive field. Although this method can extract the information of the object and its surrounding area at the same time, the size of the surrounding area is limited and the feature maps still

cannot describe the scene semantics well. Zou et al. [62] realized small-objects detection in high-resolution remote sensing images based on Bayesian priors algorithm, and optimized the memory overhead when processing large images. On the basis of RetinaNet, a refined single-stage detector with feature refinement for rotating object (R3Det) [63] predicts the rotated bounding boxes through the anchor with angles and the first head network. In addition, a feature-refinement module (FRM) is designed to reconstruct the entire feature map to solve the problem of misalignment. The FRM has a lightweight structure, rigorous and efficient code implementation, and can be easily inserted into a variety of cascaded detectors.

Different representation methods used in detectors usually have different advantages and disadvantages. Unlike methods that only use a single representation method, the proposed method introduces the bridging visual representation module (BVR) to combine multiple characterization methods. Through fusing the key-point representation algorithm and the bounding box representation algorithm, the positioning accuracy of objects is improved, while the influence of complex background is reduced.

2.2. Multi-Scale Feature Representations

For convolutional neural networks, the fusion and use of multi-scale feature maps can greatly improve the detection accuracy of the algorithm in small-target detection tasks and scenes with dramatic scale changes [3].

In the field of object detection based on convolutional neural networks, FPN [12] is one of the earliest effective ways to solve multi-scale problems, and it is also the most widely used feature pyramid structure. FPN receives multi-scale features from the backbone structure and then builds a top-down information transfer path to enhance the representation capabilities of the multi-scale features. While retaining the top-down information flow path of FPN, path aggregation network (PANet) [64] adds a bottom-up information transmission path to realize the two-way interaction between shallow features and deep features. Furthermore, the adaptively spatial feature fusion (ASFF) [65] is designed with dense connections to transfer information between features of different scales.

The scale-transfer module proposed by scale-transferrable detection network (STDN) [66] reconstructs multi-scale feature maps without introducing new parameters. Kong et al. [67] first fused multi-scale feature maps and then used a global attention branch to reconstruct these features. Both augmented feature pyramid network (AugFPN) [68] and two-level nested octave u-structure network ($U^2 - ONet$) [69] output multiple feature maps of different scales, and then perform loss calculations on each level.

Neural architecture search-based feature pyramid network (NAS-FPN) [70] adopted an adaptive search algorithm to allow the system to automatically find the optimal multi-scale feature information flow path, thereby forming a pyramid structure with a fixed connection path. This kind of network design logic is different from the common feature pyramid, which can avoid the complicated manual design process, but the automatically search process requires huge computing resources. Inspired by NAS-FPN, mobile-friendly neural architecture search-based feature pyramid network (MnasFPN) [71] added the characteristics of mobile hardware to the search algorithm. Therefore, when searching for the optimal network structure, the search algorithm not only considers accuracy as the only basis for judgment but also takes the hardware characteristics into account. Therefore, the deployment of MnasFPN on mobile is more advantageous.

Weighted bi-directional feature pyramid network (BiFPN) [72] improved multi-scale expression ability through connections across different scales and short-cut operations. OcSaFPN [3] improved the robustness of multi-scale features on noisy data by assigning different weights to feature maps from different depths.

The existing multi-scale information fusion methods mainly adopt convolution of fixed-weight parameters. By contrast, proposed method guides the multi-scale features fusion with weight parameters dynamically generated by conditional convolutions. This

variable parameter form is more flexible and can better adapt to aerial images with dramatic scale changes.

2.3. Conditional Convolution Mechanism

Conditional convolution, which can also be called dynamic filter, was first proposed by Jia et al. [73]. Different from the traditional convolutional layers with fixed-weight parameters in the inference stage, the parameters of a conditional convolutional layer are constantly changing with different input data. Therefore, the parameter form is more flexible. This variable parameter form can better adapt to the input data, thus it has gradually attracted the attention of many researchers.

At the same time, hyper network [74] is proposed to generate weights for another network. Later, this mechanism is also adopted to the style transfer task [75]. A kind of dynamic upsampling filter is used by Jo et al. [76] for the task of high-resolution video reconstruction. Similarly, magnification-arbitrary network for super-resolution (Meta-SR) also adopted the idea of dynamic parameter generation for super-resolution reconstruction tasks [77]. Conditionally parameterized convolutions (Condconv) [78] described the logic of conditional convolution in detail and used the form of group convolution to deal with the situation of multiple data in a batch. Wu et al. [79] generated optical flow data based on a dynamic filtering strategy. Both Harley et al. [80] and conditional convolutions for instance segmentation (CondInst) [81] adopted the conditional convolution mechanism to predict the instance masks. Xue et al. [82] adopted the conditional convolution mechanism to generate the future frames. Both Sagong et al. [83] and Liu et al. [84] introduced the conditional convolution mechanism into the generative adversarial network. Top-to-down lane-detection framework based on conditional convolution (CondLaneNet) [85] and full-scale fusion network based on conditional dilated convolution (ConDinet++) [86] are used for lane line extraction tasks and the road extraction tasks, respectively, in remote sensing images.

In summary, most of the current research on conditional convolution focus on pixel-level tasks, and the number of related research is generally small. Applications related to remote sensing images are even rarer.

3. Proposed Method

In this section, we introduce the RelationRS algorithm. The flow chart of the proposed object detection method is shown in Figure 2. The proposed RealationRS is based on the classic anchor-free detector, namely FCOS [22], with the backbone module (residual network-50 (ResNet50) [28]) and FPN structure [12]. Firstly, the input aerial image data flows through the backbone network and FPN module, and then feature maps of five scales can be obtained (P_2, P_3, P_4, P_5, P_6). In order to better adapt to multi-scale object detection tasks, the dual relationship module is designed to fuse features of different scales, which is explained in Section 3.1. On the one hand, this dual relationship module can learn the relationship of an object at different scales, and dynamically generate the fusion weights according to the input data to guide the fusion of multi-scale information. On the other hand, the dual relationship module can learn the potential scene semantics between different patches in one batch, and improve the detection accuracy through the comparison between different scenes. In Section 3.2, on the basis of fusion of multi-scale information, we use the bridging visual representations module to suppress the influence of complex background information in aerial images, and improve accuracy through the combination of multiple features. Aerial imagery usually adopts top-view perspective imaging; thus, there is little occlusion between objects (Figure 3). One of the disadvantages of the key-point-based object-detection algorithm is that it is not robust enough when encountering occlusion problems, and its advantage lies in better positioning accuracy. Based on the above reasons, the use of key-point detection technology in aerial imagery can achieve strengths and avoid weaknesses. By combining rectangular bounding box detection, center detection, corner detection, and classification, the interference of complex

background information can be suppressed, and the positioning accuracy of objects on complex background data can be improved. Finally, high-precision aerial image object detection can be achieved.

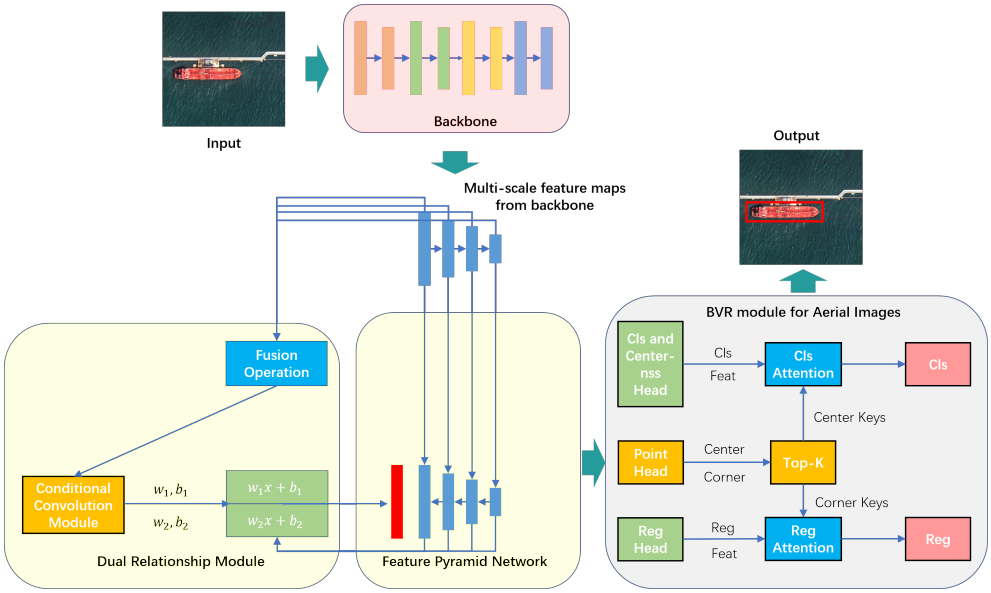

Figure 2. The proposed framework, namely, the relationship representation network for object detection in aerial images (RelationRS), is made up of three components: the baseline network, which is made up of the fully convolutional single-stage detector with no anchor setting (FCOS) [22] and the feature pyramid network (FPN) [12]; the dual relationship module, which learns the potential relationship between different scales of the objects and the potential relationship between different scenes of aerial images in one batch; and the bridging visual representation module for aerial image object detection task, which learns the potential relationship between different coordinate representations based on BVR module [31].

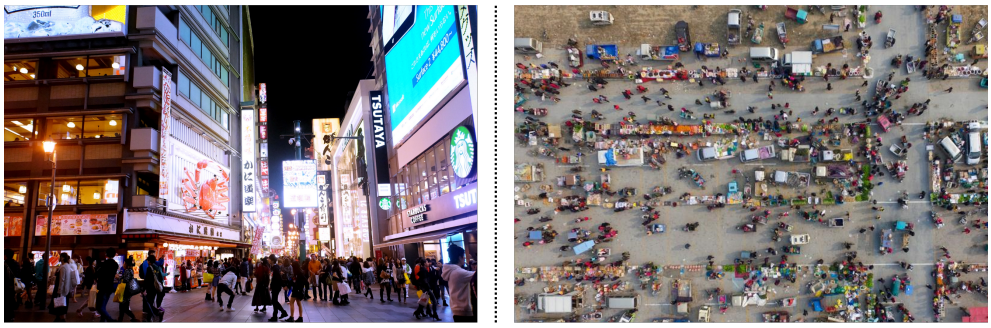

Figure 3. Comparison of target occlusion in images under different viewing angles. (**a**) There are serious occlusion problems between different people in natural images. (**b**) In images acquired by drones, due to the overhead perspective, there is almost no problem of mutual occlusion between people and vehicle objects.

3.1. Dual Relationship Module (DRM)

Aerial images have obvious characteristics of diverse scales and the existence of scene semantics. To deal with the scale changes within and between classes, FPN [12], PANet [64], NAS-FPN [70], MnasFPN [71], BiFPN [72], OcSaFPN [3], etc. have all been proposed. These methods effectively improve the multi-scale object detection problem. However, the structures and weights of these methods are fixed in the inference stage and will not change according to the input data. As shown in Figure 4, for different aerial image patches,

the semantic information of the scene contained in it is different, and the object types and scales in the two scenes are also quite different. Based on the above reasons and inspired by CondInst [81], the dual relationship module is designed to learn scale changes from multi-scale information, implicitly extract the connections and differences between the scenes contained in different patches in the one batch. In addition, the neural network parameters of the multi-scale information fusion module are dynamically generated to guide the fusion of multi-scale features with semantic information of the input data.

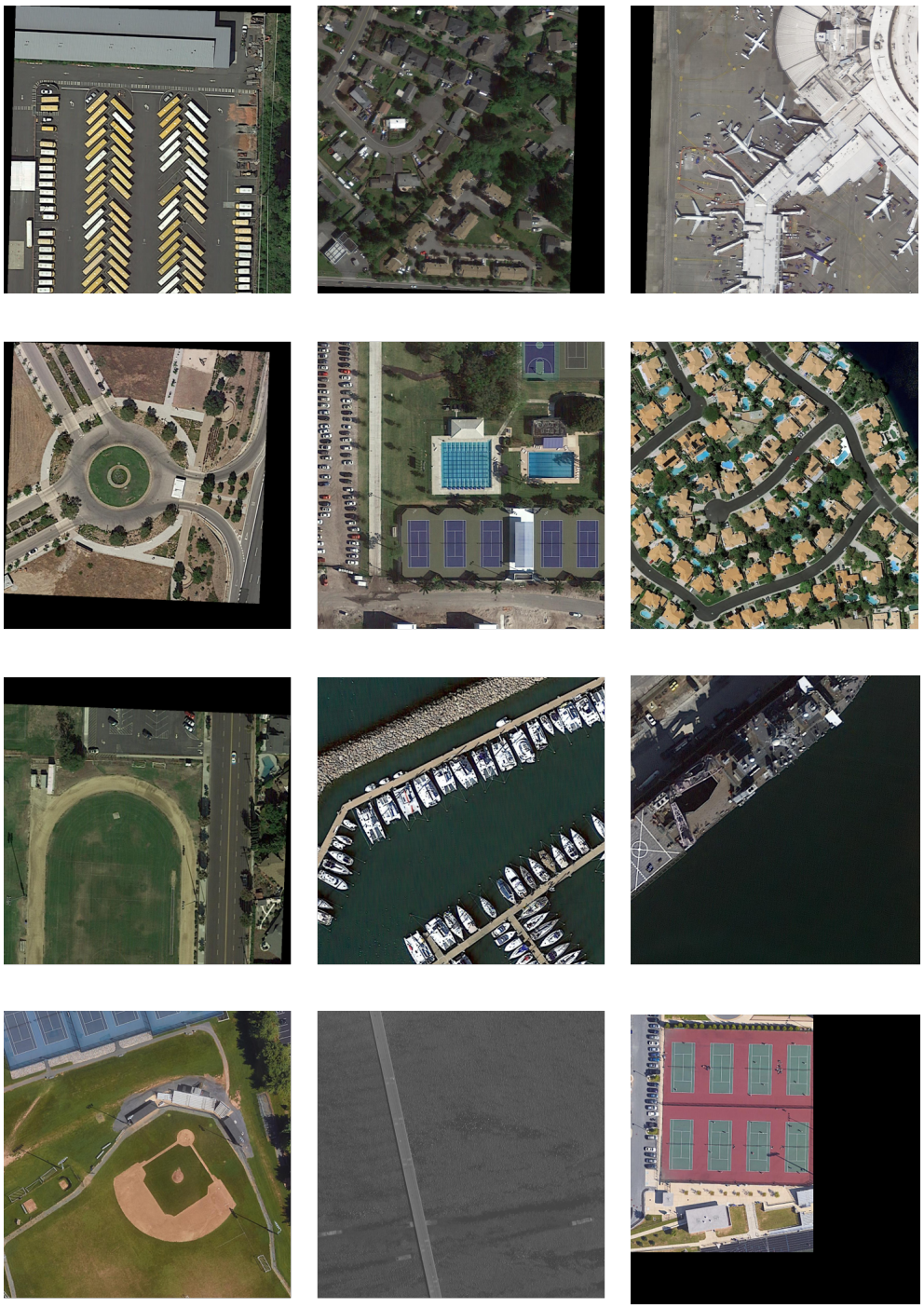

Figure 4. Sample aerial image patches with different scenes. The sizes of patches are all 1024 × 1024. Different patches have different scene semantics, forming a potential semantic contrast with each other. There are also intra-class and inter-class scale differences for objects between different scenarios.

Figure 5 shows the construction of the dual relationship module. Taking $batchsize = 2$ as an example, the four scale feature maps extracted from the backbone network are marked as C_2, C_3, C_4, C_5. Then, after taking C_2, C_3, C_4, C_5 as input, the feature maps output by FPN [12] can be marked as P_2, P_3, P_4, P_5, P_6. Among them, P_6 is generated by P_5 with the maximum pooling operation. Inspired by CondInst, the key point of the dual relationship module is the generation of P_2' feature maps, and the generation process of P_2' can be described by Equation (1):

$$P_2' = Conv2(Conv1(concate(resize((P_2, P_3, P_4, P_5))))), \quad (1)$$

where $Conv1(\cdot)$ and $Conv2(\cdot)$ denote two convolution operations with kernel size [1, 1], $concate(\cdot)$ denotes the concatenation. The parameters of $Conv1(\cdot)$ and $Conv2(\cdot)$ are obtained by CondConv [78], and the process can be expressed by Equation (2):

$$w_1, w_2, b_1, b_2 = CondConv(FO(C_2, C_3, C_4, C_5)), \quad (2)$$

where $CondConv(\cdot)$ denotes the conditional convolution module from CondConv [78], $FO(\cdot)$ denotes the fusion operation seen in Figure 5 and can be built by Equation (3):

$$FO(C_2, C_3, C_4, C_5) = Conv_{3\times3}(concate(resize(C_2, C_3, C_4, C_5))), \quad (3)$$

where $Conv_{3\times3}(\cdot)$ denotes the convolution operation with kernel size [3, 3].

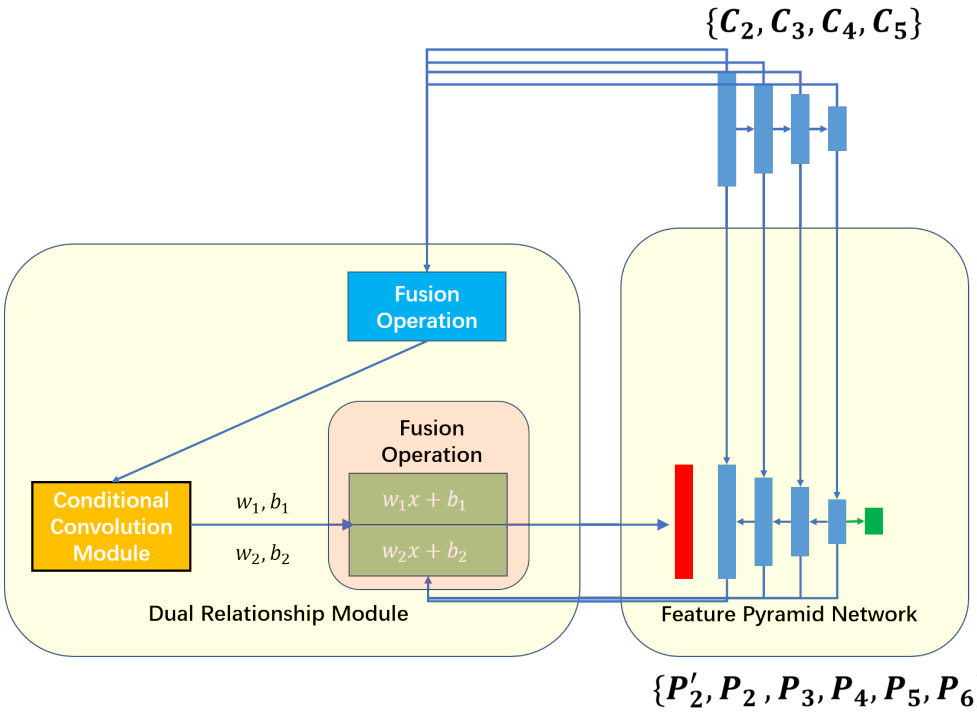

Figure 5. The construction process for the dual relationship module.

Based on the above formulas, C_2, C_3, C_4, C_5 are resized and concatenated in the channel dimension. In order to obtain w_1, w_2, b_1, b_2, the feature map after fusion first passes through a convolutional layer with kernel size [3, 3] for feature alignment and channel dimensionality reduction. Then, the dimensionality-reduced feature map is sent to CondConv module to generate the required weights and bias values. To solve the situation where batchsize is not equal to 1, we treat different patches in a batch as different experts. This is different from the way in CondInst. Finally, we generate the parameters to initialize the two convolutional layers for the fusion of P_2, P_3, P_4, P_5, and finally output the required feature maps P_2'. In this way, we obtain a series of multi-scale feature maps $P_2', P_2, P_3, P_4, P_5, P_6$ and

send them to the head network. To reduce the amount of parameters (w_1, w_2, b_1, b_2), the group convolution mechanism from AlexNet [87] is used in $Conv1(\cdot)$ and $Conv2(\cdot)$.

It is worth noting that we treat two patches in a batch as experts (Figure 6), and then generate weight parameters. These parameters potentially obtain the relationships and differences between two scenarios; thus, the network can dynamically extract features based on the input data. This is different from the idea of using semantic extraction branches in CAD-Net [30] and GLS-Net [2] to extract the semantics of a single patch scene.

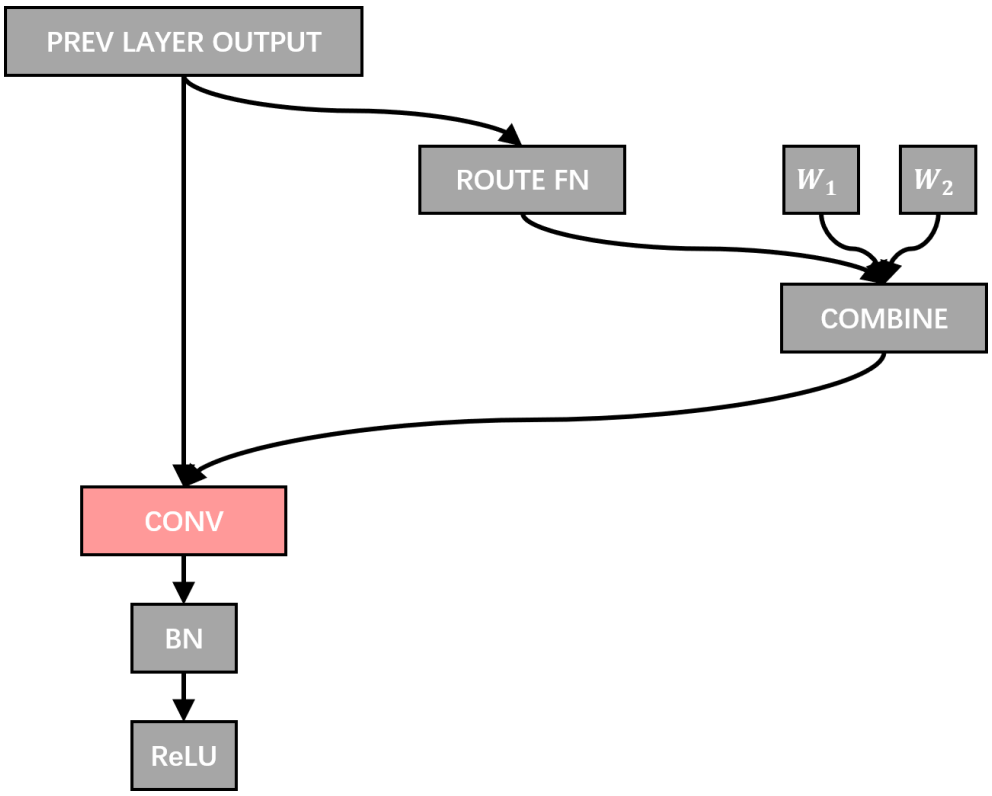

Figure 6. CondConv [78] layer architecture with $n = 2$ kernels. CondConv: $(\alpha_1 W_1 + \alpha_2 W_2) \times x$.

3.2. Bridging Visual Representations for Object Detection in Aerial Images

Generally speaking, different representation methods usually lead detectors to perform well in different aspects. According to the structures of the current object detection networks, the two-stage detector can usually obtain a more accurate object category prediction. The detection method based on the center point can improve the detection accuracy of small objects. The corner-based method reduces the characterization dimension of the bounding boxes; thus, it has more advantages in positioning tasks [31]. As shown in Figure 3, in the face of the character that objects in aerial images are less occluded, we believe that the introduction of the key-point detection algorithm based on FCOS can suppress the influence of complex backgrounds and improve detection accuracy. Based on this point of view, the BVR module is introduced into the aerial image object detection task. Through the combination of multiple characterization methods, the accuracy of aerial image detection tasks for complex backgrounds can be improved.

For a detector, the main idea of the BVR module is to regard its main representation as the master representation. Thus, other auxiliary representations are adopts to enhance the master representation by a kind of transformer module [88]. That is, the transformer mechanism is used to bridge different visual representations.

As for the anchor-free algorithm (FOCS), the center point location and the corresponding features are regarded as the master representation and the query input at the same time. Compared with standard FOCS head network, the BVR module constructs an additional point head network (Figure 7). The point head network consists of two shared convolu-

tional layers with kernel size [3, 3], followed by two independent sub-networks to predict the scores and sub-pixel offsets for center and corner prediction [31]. The representations produced by the point head network are regarded as the auxiliary representation and the keys in the transformer algorithm. To reduce the amount of calculation, a top-k key selection strategy is adopted to control the set of keys no larger than k (default = 50), according to their corner-ness scores. Indeed, the cosine/sine location embedding algorithm is used to reduce the complexity of coordinate representations. Here, according to the characteristic of the aerial images, the maximum number of key points is set to 400.

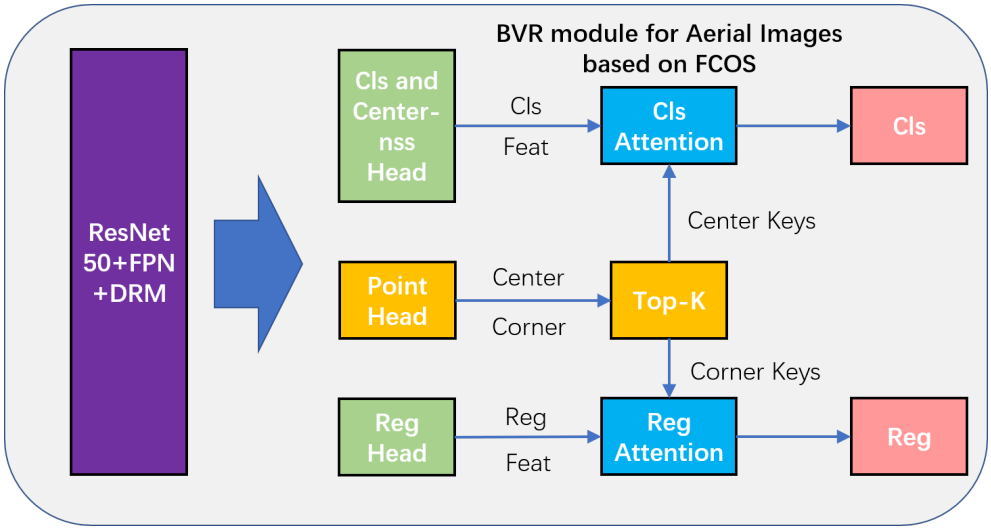

Figure 7. The construction process for the BVR module based on FCOS.

Based on the above settings of the queries and the keys, the enhanced features $f_i'^q$ can be calcuted by Equation (4):

$$f_i'^q = f_i^q + \sum_j S(f_i^q, f_j^k, g_i^q, g_j^k) \cdot T_v(f_j^k), \tag{4}$$

where f_i^q and g_i^q are the input feature and geometric vector for a *query* instance i; f_j^k and g_j^k are the input feature and geometric vector for a *key* instance j; $T_v(\cdot)$ is a linear *value* transformation function; $S(\cdot)$ is a similarity function between i and j [31]. $S(\cdot)$ can be described as Equation (5):

$$S(f_i^q, f_j^k, g_i^q, g_j^k) = softmax_j(S^A(f_i^q, f_j^k) + S^G(g_i^q, g_j^k)), \tag{5}$$

where $S^A(f_i^q, f_j^k)$ denotes the appearance similarity computed by a scaled dot product between *query* and *key* features, and $S^G(g_i^q, g_j^k)$ denotes a geometric term computed by cosine/sine location embedding-based method [31].

The BVR module based on FCOS can be seen in Figure 7.

4. Experiments and Results

For proving the effectiveness of the proposed method, a widely used "A Large-Scale Dataset for Object Detection in Aerial images" (DOTA) [1] dataset was used in the experiments for the object detection task in aerial images. In this chapter, the DOTA1.0 dataset, the implementation details, and the ablation studies conducted with the proposed method can be introduced in detail in order.

4.1. Dataset

DOTA1.0. DOTA1.0 dataset is one of the largest published open-access datasets for object detection in aerial images. The dataset consists of 2806 large-size aerial images from

Google Earth and satellites including Julang-1 (LJ-1) and the Gaofen-2 satellite (GF-2), with 188,282 annotated bounding boxes in 15 categories, including plane, baseball diamond (BD), bridge, ground track field (GTF), small vehicle (SV), large vehicle (LV), ship, tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball field (SBF), roundabout (RA), harbor, swimming pool (SP), and helicopter (HC). Most of the ground sampling distance(GSD) values of the images in the DOTA1.0 dataset are better than 1 m. The DOTA1.0 dataset has the characteristics of diverse scenarios, categories, scales, etc. It is still challenging to achieve high-precision object detection with this dataset.

In the experiments, the training set and validation set of DOTA1.0 were used in the training stage, and the test set was used for the inference stage and evaluation stage. The original images were all cropped to a size of 1024×1024, overlapping by 500 pixels. If the size of a patch is less than 1024×1024, the zero padding method was adopted for completion. Based on the above settings, we obtained a total of 38,504 patches for the training stage and a total of 20,012 patches for the evaluation task. Since the ground truth files of the test set from the DOTA1.0 dataset were not disclosed, we submitted the final test results to the online evaluation website in the format of '.txt' (https://captain-whu.github.io/DOTA/evaluation.html (accessed on 6 July 2021)).

4.2. Evaluation Metrics

For quantitative accuracy evaluation, the mean average precision (mAP) is used in this paper. The mAP describes the mean value of the average precision (AP) values for multiple categories in a dataset. For a certain category, the AP value is the area enclosed by the coordinate axis and the broken line in the corresponding precision-recall graph. The larger the area, the higher its AP value. The details of the evaluation follow the official DOTA1.0 evaluation website (https://captain-whu.github.io/DOTA/evaluation.html (accessed on 6 July 2021)).

4.3. Implementation Details

For the realization of the RelationRS, we built the baseline network based on FCOS [22] with FPN [12]. The ResNet50 [28] pretrained on ImageNet [89] was adopted as the backbone network. A series of experiments were designed to better evaluate the effects of the dual relationship module and the bridging visual representations module for aerial image object detection in this paper. The environment used was a single NVIDIA Tesla V100 GPU with 16 GB memory, along with the PyTorch 1.8.0 and Python 3.7.10 deep learning frameworks. The initial learning rate was 0.0025, the batch size of the input data was 2, the value of the momentum is 0.9, the value of the weight decay was 0.0001, and the minibatch stochastic gradient descent (SGD) was also used for optimization. And the project was build on the mmdetection v2.7.0 [90].

4.4. Ablation Experiments

Two ablation experiments were used to further discuss the influence of the dual relationship module and the bridging visual representations module. Here, the abbreviations in the DOTA data set are explained again: plane, baseball diamond (BD), bridge, ground track field (GTF), small vehicle (SV), large vehicle (LV), ship, tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball field (SBF), roundabout (RA), harbor, swimming pool (SP), and helicopter (HC).

4.4.1. Dual Relationship Module

We conducted the ablation experiment to verify the effectiveness of the proposed dual relationship module. The baseline is the FCOS algorithm. +DRM means the combination of the FCOS and the dual relationship module. The difference between the baseline and the +DRM is only whether the dual relationship module is additionally used, and the parameters used in the experiment are strictly kept consistent.

From Table 1, +DRM obtains a mAP of 65.63%, which is 1.38% higher than the mAP value of the baseline (64.25%). For DOTA datasets with 15 categories, the baseline method

only has advantages in three categories, plane, soccer-ball field (SBF), and basketball court (BC), which are 0.17%, 4.21% and 5.4% higher than the values of +DRM. This indicates that the performance of +DRM is not stable enough for objects with relatively large scales. For small objects, +DRM has achieved better accuracy in multiple categories. The AP values of small vehicle (SV), large vehicle (LV), ship, storage tank (ST), and helicopter (HC) of the +DRM outperform the values of the baseline by 1.66%, 3.07%, 4.97%, 6.48%, and 6.46%. In addition, The AP values of baseball diamond (BD), bridge, ground track field (GTF), roundabout (RA), harbor and swimming pool (SP) of the +DRM are also higher than values of the baseline. Therefore, the +DRM method can effectively improve the detection accuracy of small targets.

Table 1. Detection accuracy in the ablation study of using DRM or not with DOTA test dataset. The bold numbers denote the highest values in each class.

Method	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	mAP(%)
	BC	ST	SBF	RA	Harbor	SP	HC		
baseline	**88.12**	70.77	44.04	47.46	76.36	65.34	77.96	90.83	64.25
	74.31	78.37	**48.3**	52.62	72.25	42.77	34.27		
+DRM	87.95	**71.66**	**44.1**	**52.48**	**78.02**	**68.41**	**82.93**	90.83	**65.63**
	68.91	**84.85**	44.09	**53.6**	**72.44**	**43.42**	**40.73**		

4.4.2. Bridging Visual Representations for Object Detection in Aerial Images

To evaluate the efficiency of the bridging visual representations module in aerial images, an ablation experiment are designed to compare the BVR module with the baseline in the DOTA test dataset. From Table 2, the baseline is the FCOS algorithm the same as Section 4.4.1. +BVR is a combination of the baseline method and the bridging visual representations module.

For the 10 categories—plane, baseball diamond (BD), bridge, ground track field (GTF), small vehicle (SV), large vehicle (LV), ship, storage tank (ST), swimming pool (SP) and helicopter (HC)—the AP values of the +BVR are higher than values of the baseline by 1.2%, 1.98%, 1.17%, 5.55%, 2%, 0.31%, 0.84%, 1.9%, 5.01%, and 9.8%, respectively. Thus, +BVR achieves a higher mAP values by 1.67%. The value increase of the mAP proves the effectiveness of the BVR method in the field of aerial image detection.

Table 2. Detection accuracy in the ablation study of using BVR module or not with DOTA test dataset. The bold numbers denote the highest values in each class.

Method	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	mAP(%)
	BC	ST	SBF	RA	Harbor	SP	HC		
baseline	88.12	70.77	44.04	47.46	76.36	65.34	77.96	90.83	64.25
	74.31	78.37	**48.3**	**52.62**	**72.25**	42.77	34.27		
+BVR	**89.32**	**72.75**	**45.21**	**53.01**	**78.36**	**65.65**	**78.74**	90.83	**65.92**
	72.88	**80.27**	45.42	52.36	72.12	**47.78**	**44.07**		

4.5. Comparison with the State-of-the-Art

To examine and evaluate the performance of the proposed framework RelationRS, the proposed framework is compared with the state-of-the-art algorithms with the DOTA test dataset. Table 3 shows the AP and the mAP values of different algorithms.

Table 3. Comparisons with the state-of-the-art single-stage-based detectors in the DOTA1.0 test dataset with horizontal bounding boxes. The baseline is the FCOS algorithm. +DRM means the combination of the FCOS and the dual relationship module. +BVR is a combination of the baseline method and the bridging visual representations module, and the RelationRS is a network adding the DRM and the BVR module to the baseline detector. The red numbers and the blue numbers denote the highest values and the second highest values in each class.

Method	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	mAP(%)
	BC	ST	SBF	RA	Harbor	SP	HC		
YOLOv3-tiny [18]	61.48	24.35	4.3	15.49	20.27	30.22	26.96	72	25.73
	26.21	22.91	14.05	7.27	28.78	27.07	4.55		
SSD [21]	57.85	32.79	16.14	18.67	0.05	36.93	24.74	81.16	29.86
	25.1	47.47	11.22	31.53	14.12	9.09	0		
YOLOv2 [17]	76.9	33.87	22.73	34.88	38.73	32.02	52.37	61.65	39.2
	48.54	33.91	29.27	36.83	36.44	38.26	11.61		
RetinaNet [23]	78.22	53.41	26.38	42.27	63.64	52.63	73.19	87.17	50.39
	44.64	57.99	18.03	51	43.39	56.56	7.44		
YOLOv3 [18]	79	77.1	33.9	68.1	52.8	52.2	49.8	89.9	60
	74.8	59.2	55.5	49	61.5	55.9	41.7		
SBL [91]	89.15	66.04	46.79	52.56	73.06	66.13	78.66	90.85	64.77
	67.4	72.22	39.88	56.89	69.58	67.73	34.74		
$SFFM^d$ [92]	88.1	82.4	47.7	72.9	45.9	73.5	64.4	90.4	66.3
	66.7	50.1	54	60.1	77.8	51.7	69.5		
baseline [22]	88.12	70.77	44.04	47.46	76.36	65.34	77.96	90.83	64.25
	74.31	78.37	48.3	52.62	72.25	42.77	34.27		
+DRM	87.95	71.66	44.1	52.48	78.02	68.41	82.93	90.83	65.63
	68.91	84.85	44.09	53.6	72.44	43.42	40.73		
+BVR	89.32	72.75	45.21	53.01	78.36	65.65	78.74	90.83	65.92
	72.88	80.27	45.42	52.36	72.12	47.78	44.07		
RelationRS	88.27	72.96	45.47	53.7	79.73	70.98	82.38	90.83	66.81
	69.86	83.29	45.26	54.61	72.79	47.85	44.18		

As shown in Table 3, the proposed RelationRS achieves the highest mAP value, and it outperfors YOLOv3-tiny [18], SSD [21], YOLOv2 [17], RetinaNet [23], YOLOv3 [18], SBL [91], $SFFM^d$ [92], and FCOS (baseline) [22] by 41.08%, 36.95%, 27.61%, 16.42%, 6.81%, 2.04%, and 0.51%. For 15 categories of objects, RelationRS obtains the highest mAP value on 1 class (small vehicle) and the second highest mAP value on 6 classes (large vehicle, ship, tennis court, storage tank, harbor, and helicopter).

RelationRS has good detection performance for small objects, such as small vehicle (SV), large vehicle (LV), ship, storage tank (ST), and helicopter (HC), etc. This phenomenon is similar to the one discussed in Section 4.4.1. In addition, the BVR module is also used in the RelationRS method to improve the performance of the detector on images with complex background. Therefore, the overall accuracy of multiple categories is improved to a certain extent. Interestingly, $SFFM^d$ obtains the highest mAP value on 7 classes and the second highest mAP value on 1 classes, but its mAP value is slightly lower than the RelationRS method.

As shown in Figure 8, the first line is the visualization of the baseline inference result, the second line is the visualization of the proposed RelationRS algorithm in this paper. For the first column, it can be seen that the baseline algorithm missed a small-vehicle target, while the RelationRS correctly detected the target. The overall probability score of the vehicle target in the picture has been improved, which is consistent with the numerical performance in Table 3. This can prove the effectiveness of the RelationRS algorithm for multi-scale feature fusion. Similarly, a vehicle in the central area in (b) is falsely detected as a ship with a probability of 0.53, while in (f) the target is still falsely detected, but its probability drops to 0.33. The introduction of scene information (from DRM) and accuracy position representation (from BVR) can suppress false detection results that violate logic to a certain extent. Compared with (b), the overall probability scores of vehicle objects in (f) are improved. For the third column, the false detection at *NO*.4 in (c) is eliminated in (g), and it is possible to avoid the detection of objects on the coast as ships based on the scene semantics. It is very interesting that the false detection at *NO*.2 in (c) can be eliminated by the combination of the two representations of BVR, but the scene information of the complex area will affect the detection of difficult targets, such as the miss detection at the *NO*.2 in (g). In the fourth column, BVR suppresses the background and improves the positioning accuracy of the target. Combined with the scene information obtained by DRM, it is easy to eliminate the false detection at *NO*.1 in (d), hence the better detection results obtained in (h).

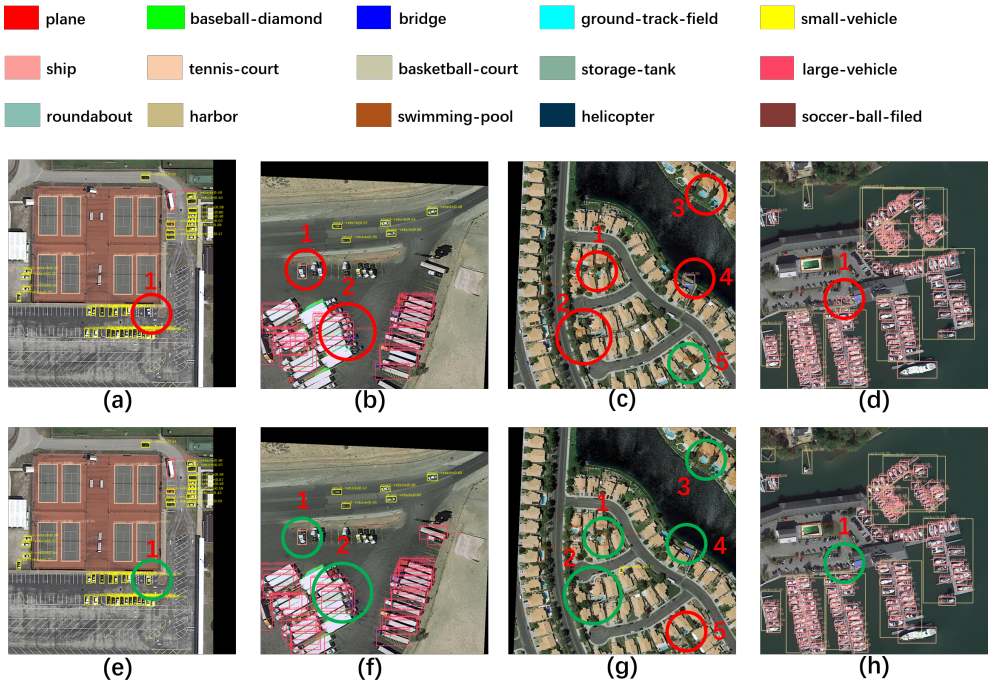

Figure 8. The visualization comparison between the baseline method and the proposed RelationRS in this paper. The first line (**a**–**d**) is the visualization of the baseline inference result, and the second line (**e**–**h**) is the visualization of the RelationRS inference result. Each column represents the detection results of two algorithms in the same scene, where the red circle represents the failure case in the column, and the green circle represents the better case in the column. Circles with the same number in the same column form a contrasting pair.

Table 4 shows the inference speed and memory usage of the proposed method. Although the RealationRS algorithm proposed in this paper has certain disadvantages in terms of speed and GPU memory usage, it is still a useful exploration for the extraction and use of remote sensing image scene semantics and the combined use of multiple representation methods.

Table 4. The inference speed (frames per second, FPS) and graphics processing unit memory occupancy of the method proposed in this paper. Bold numbers indicate the best value in that column.

Method	FPS (fps)	Memory (MByte)
baseline	**11.65**	**1691**
+BVR	7.97	2081
RelationRS	7.67	2157

In conclusion, RelationRS constructs a novel dual relationship module to guide the fusion of multi-scale features. DRM learns the comparison information of scenes from different patches in one batch and dynamically generates the weight parameters of multi-scale fusion through conditional convolution. Furthermore, the bridging visual representations module for natural image object detection is introduced into RelationRS to improve the performance of the detector on aerial images with complex background information. To the best of our knowledge, this is the first time that BVR module is introduced into remote sensing image object detection task. The BVR module can combine the two representation methods of rectangular box and key-point representation to better locate the target position in the complex background, so that the network can better learn the information of the target itself. Some examples of detection results in different scenarios are shown in Figure 9.

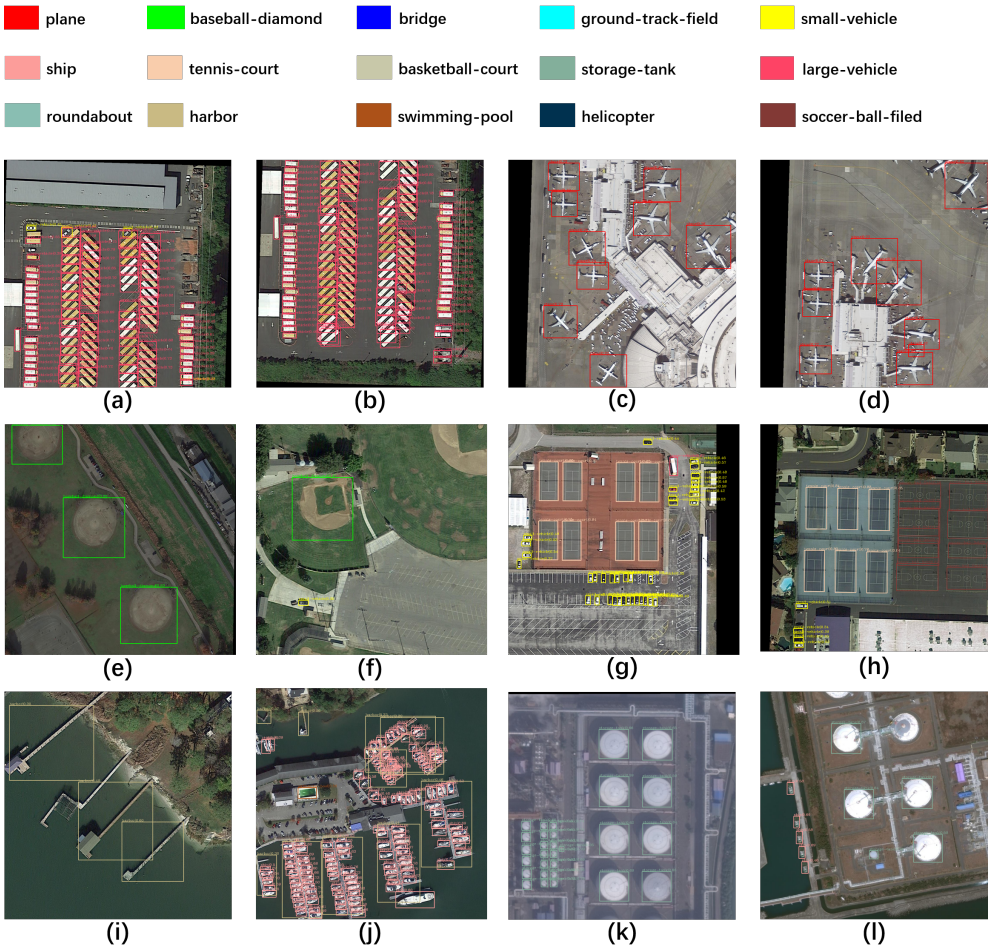

Figure 9. Object detection results on DOTA1.0 dataset. (**a**–**l**) are schematic pictures of detection results in different scenarios.

5. Conclusions

In this paper, a single-stage-based object detector for aerial images is proposed, namely RelationRS. The framework combines a dual relationship module and a bridging visual

representations module to solve the problem of multi-scale fusion and improve the object detection accuracy in aerial images with complex backgrounds. The proposed novel dual relationship module in this paper can extract the relationship between different scenes according to the input images and dynamically generate the weight parameters required for feature maps fusion. This feature-fusion method can improve the detection accuracy and probability score of small objects. In addition, based on the characteristics of less mutual occlusion of objects in remote sensing images, from the viewpoint of introducing a key-point detection algorithm, we prove the effectiveness of the bridging visual representations module in the field of aerial image object detection. The BVR module can combine the two representation methods of the rectangular box and the key-point representation to better locate the target position in the complex background. To the best of our knowledge, this is the first demonstration of the performance of the BVR module on remote sensing images. Furthermore, BVR-based accurate positioning and DRM-based scene information introduction can eliminate false detections with serious scene logic errors to a certain extent, such as vehicles onshore being detected as ships. The experiments under taken with the public DOTA1.0 dataset confirmed the remarkable performance of the proposed method.

On the other hand, single-stage detectors are still not as accurate as two-stage detectors. The current neural networks still cannot be better explained. Therefore, how to better explain the features extracted by neural networks and combine the imaging parameters of aerial images is one of the key points to improve the detection accuracy of aerial image in the future.

Author Contributions: Conceptualization, Z.L. and X.Z.; methodology, Z.L. and C.L.; software, B.L.; validation, H.W.; formal analysis, C.S.; resource, P.H.; data curation, Q.L.; writing—original draft preparation, Y.L.; writing—review and editing, X.Z.; visualization, B.L.; supervision, J.X.; project administration, Z.L.; funding acquisition, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Major Science and Technology Projects of China grant number (30-H32A01-9005-13/15).

Data Availability Statement: The author declares that the experimental dataset can be obtained from the DOTA dataset offical website (https://captain-whu.github.io/DOTA/, accessed on 6 July 2021).

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

References

1. Xia, G.S.; Bai, X.; Zhang, L.P.; Serge, B.; Marcello, P. 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, 18–22 June 2018; pp. 7132–7141.
2. Li, C.; Luo, B.; Hong, H.; Su, X.; Wang, Y.; Liu, J.; Wang, C.; Zhang, J.; Wei, L. Object Detection Based on Global-Local Saliency Constraint in Aerial Images. *Remote Sens.* **2020**, *12*, 1435. [CrossRef]
3. Li, C.; Liu, J.; Hong, H.; Mao, W.; Wang, C.; Hu, C.; Su, X.; Luo, B. Object Detection based on OcSaFPN in Aerial Images with Noise. *arXiv* **2020**, arXiv:2012.09859.
4. Huyan, L.; Bai, Y.; Li, Y.; Jiang, D.; Zhang, Y.; Zhou, Q.; Wei, J.; Liu, J.; Zhang, Y.; Cui, T. A Lightweight Object Detection Framework for Remote Sensing Images. *Remote Sens.* **2021**, *13*, 683. [CrossRef]
5. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2017**, *13*, 1074–1078. [CrossRef]
6. Yang, F.; Xu, Q.; Li, B. Ship Detection From Optical Satellite Images Based on Saliency Segmentation and Structure-LBP Feature. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 602–606. [CrossRef]
7. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
8. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef]
9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

10. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
12. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
13. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
15. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; Lecun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–26 April 2014.
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 779–788.
17. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
18. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
19. Bochkovskiy, A.; Wang, C.; Liao, H. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
20. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C. Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
22. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9627–9636.
23. 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.
24. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 734–750.
25. Law, H.; Teng, Y.; Russakovsky, O.; Deng, J. CornerNet-Lite: Efficient Keypoint Based Object Detection. *arXiv* **2019**, arXiv:1904.08900.
26. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Shi, J. FoveaBox: Beyond Anchor-based Object Detector. *arXiv* **2019**, arXiv:1904.03797.
27. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 770–778.
29. Dong, Y.; Chen, F.; Han, S.; Liu, H. Ship Object Detection of Remote Sensing Image Based on Visual Attention. *Remote Sens.* **2021**, *13*, 3192. [CrossRef]
30. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
31. Cheng, C.; Wei, F.; Hu, H. Relationnet++: Bridging visual representations for object detection via transformer decoder. *arXiv* **2020**, arXiv:2010.15831.
32. Zou, Z.; Shi, Z. Ship detection in spaceborne optical image with SVD networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *54*, 5832–5845. [CrossRef]
33. Van de Sande, K.E.; Uijlings, J.R.; Gevers, T.; Smeulders, A.W. Segmentation as selective search for object recognition. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–11 November 2011; p. 7.
34. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A. W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
35. Dong, R.; Xu, D.; Zhao, J.; Jiao, L.; An, J. Sig-NMS-Based Faster R-CNN Combining Transfer Learning for Small Target Detection in VHR Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8534–8545. [CrossRef]
36. 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]
37. Xiao, Z.; Gong, Y.; Long, Y.; Li, D.; Wang, X.; Liu, H. Airport detection based on a multiscale fusion feature for optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1469–1473. [CrossRef]
38. Xu, Z.; Xu, X.; Wang, L.; Yang, R.; Pu, F. Deformable convnet with aspect ratio constrained nms for object detection in remote sensing imagery. *Remote Sens.* **2017**, *9*, 1312. [CrossRef]
39. Ren, Y.; Zhu, C.; Xiao, S. Small object detection in optical remote sensing images via modified faster R-CNN. *Appl. Sci.* **2018**, *8*, 813. [CrossRef]
40. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.

41. Hamaguchi, R.; Fujita, A.; Nemoto, K.; Imaizumi, T.; Hikosaka, S. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1442–1450.
42. Yu, F.; Vladlen, K. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
43. Wu, Y.; Zhang, K.; Wang, J.; Wang, Y.; Wang, Q.; Li, Q. CDD-Net: A Context-Driven Detection Network for Multiclass Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
44. Qiu, H.; Li, H.; Wu, Q.; Meng, F.; Ngan, K.N.; Shi, H. A2RMNet: Adaptively Aspect Ratio Multi-Scale Network for Object Detection in Remote Sensing Images. *Remote Sens.* **2019**, *11*, 1594. [CrossRef]
45. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. *arXiv* **2017**, arXiv:1706.09579.
46. Ma, J.Q.; Shao, W.Y.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.B.; Xue, X.Y. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimedia* **2018**, *20*, 3111–3122. [CrossRef]
47. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2337–2348. [CrossRef]
48. Ding, J.; Xue, N.; Long, Y.; Xia, G.X.; Lu, Q.K. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2844–2853.
49. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8232–8241.
50. Fang, Y.; Ding, G.; Li, J.; Fang, Z. Deep3DSaliency: Deep Stereoscopic Video Saliency Detection Model by 3D Convolutional Networks. *IEEE Trans. Image Process.* **2019**, *28*, 2305–2318. [CrossRef] [PubMed]
51. Jian, M.; Wang, J.; Yu, H.; Wang, G.; Meng, X.; Yang, L.; Dong, J.; Yin, Y. Visual saliency detection by integrating spatial position prior of object with background cues. *Expert Syst. Appl.* **2021**, *168*, 114219. [CrossRef]
52. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R. RADet: Refine Feature Pyramid Network and Multi-Layer Attention Network for Arbitrary-Oriented Object Detection of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 389. [CrossRef]
53. Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; Yang, W. Mask OBB: A Semantic Attention-Based Mask Oriented Bounding Box Representation for Multi-Category Object Detection in Aerial Images. *Remote Sens.* **2019**, *11*, 2930. [CrossRef]
54. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
55. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Jie, Z.; Zhang, T.; Yang, J. Learning object-wise semantic representation for detection in remote sensing imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 20–27.
56. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef]
57. Zhu, Y.; Du, J.; Wu, X. Adaptive period embedding for representing oriented objects in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7247–7257. [CrossRef]
58. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
59. Han, J.; Ding, J.; Xue, N.; Xia, G. Redet: A rotation-equivariant detector for aerial object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2786–2795.
60. Van Etten, A. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv* **2018**, arXiv:1805.09512.
61. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3377–3390. [CrossRef]
62. Zou, Z.; Shi, Z. Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. *IEEE Trans. Image Process.* **2017**, *27*, 1100–1111. [CrossRef]
63. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv* **2019**, arXiv:1908.05612.
64. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
65. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516.
66. Zhou, P.; Ni, B.; Geng, C.; Hu, J.; Xu, Y. Scale-Transferrable Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 528–537.
67. Kong, T.; Sun, F.; Tan, C.; Liu, H.; Huang, W. Deep feature pyramid reconfiguration for object detection. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 169–185.
68. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 14–29 June 2020; pp. 12595–12604.
69. Wang, C.; Li, C.; Liu, J.; Luo, B.; Su, X.; Wang, Y.; Gao, Y. U2-ONet: A Two-Level Nested Octave U-Structure Network with a Multi-Scale Attention Mechanism for Moving Object Segmentation. *Remote Sens.* **2021**, *13*, 60. [CrossRef]

70. Ghiasi, G.; Lin, T.; Le, Q. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7036–7045.
71. Chen, B.; Ghiasi, G.; Liu, H.; Lin, T.; Kalenichenko, D.; Adam, H.; Le, Q. Mnasfpn: Learning latency-aware pyramid architecture for object detection on mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 14–29 June 2020; pp. 13607–13616.
72. Tan, M.; Pang, R.; Le, Q. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 14–29 June 2020; pp. 10781–10790.
73. Jia, X.; De Brabandere, B.; Tuytelaars, T.; Gool, L. Dynamic filter networks. *Adv. Neural Inf. Process Syst.* **2016**, *29*, 667–675.
74. Ha, D.; Dai, A.; Le, Q. Hypernetworks. *arXiv* **2016**, arXiv:1609.09106.
75. Shen, F.; Yan, S.; Zeng, G. Neural style transfer via meta networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8061–8069.
76. Jo, Y.; Oh, S.W.; Kang, J.; Kim, S. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3224–3232.
77. Hu, X.; Mu, H.; Zhang, X.; Wang, Z.; Tan, T.; Sun, J. Meta-SR: A magnification-arbitrary network for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1575–1584.
78. Yang, B.; Bender, G.; Le, Q.; Ngiam, J. Condconv: Conditionally parameterized convolutions for efficient inference. *arXiv* **2019**, arXiv:1904.04971.
79. Wu, J.; Li, D.; Yang, Y.; Bajaj, C.; Ji, X. Dynamic filtering with large sampling field for convnets. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 185–200.
80. Harley, A.; Derpanis, K.; Kokkinos, I. Segmentation-aware convolutional networks using local attention masks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5038–5047.
81. Tian, Z.; Shen, C.; Chen, H. Conditional convolutions for instance segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2018; pp. 282–298.
82. Xue, T.; Wu, J.; Bouman, K.L.; Freeman, W. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. *arXiv* **2016**, arXiv:1607.02586.
83. Sagong, M.; Shin, Y.; Yeo, Y.; Park, S.; Ko, S. cGANs with Conditional Convolution Layer. *arXiv* **2019**, arXiv:1906.00709.
84. Liu, X.; Yin, G.; Shao, J.; Wang, X.; Li, H. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *arXiv* **2019**, arXiv:1910.06809.
85. Liu, L.; Chen, X.; Zhu, S.; Tan, P. CondLaneNet: A Top-to-down Lane Detection Framework Based on Conditional Convolution. *arXiv* **2021**, arXiv:2105.05003.
86. Yang, K.; Yi, J.; Chen, A.; Liu, J.; Chen, W. ConDinet++: Full-Scale Fusion Network Based on Conditional Dilated Convolution to Extract Roads From Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
87. Sutskever, I.; Hinton, G.E.; Krizhevsky, A. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *60*, 1097–1105.
88. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
89. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the International Conferenceon Neural Information Processing Systems, Lake Tahoe, ND, USA, 3–8 December 2012; pp. 1097–1105.
90. 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.
91. Sun, P.; Chen, G.; Luke, G.; Shang, Y. Salience biased loss for object detection in aerial images. *arXiv* **2018**, arXiv:1810.08103.
92. Wang, P.; Sun, X.; Diao, W.; Fu, K. Mergenet: Feature-merged network for multi-scale object detection in remote sensing images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28–2 July–August 2019; pp. 238–241.