Tiny Object Detection Enhancement for Large-Scale Remote Sensing Imagery

Tianwei Zhang* Lianru Gao† Xu Sun† and Lina Zhuang†
* Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
E-mail: zhangtianwei20@mails.ucas.ac.cn
† Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
E-mail: gaolr@aircas.ac.cn, sunxu@aircas.ac.cn, zhuangln@aircas.ac.cn

***Abstract*—The effective detection of targets of interest in large-scale remote sensing imagery has always been a crucial application task. Detecting tiny objects in these scenes poses particular challenges, as it can lead to the loss of robust features and difficulties in generating positive sample matches. To address this application scenario, this paper proposes a method that embeds the use of vague search (VGS) to match high-potential candidate regions within a coarse-to-fine detection framework. Additionally, a new centroid distance measurement (CDM) is proposed to distinguish between positive and negative proposals. The combination of VGS and CDM can effectively increases the number of positive samples that can be refined in the high-dimensional detection layer, thereby improving the precision of detecting tiny objects in large-scale scenes. We tested and validated our model on a high-resolution optical remote sensing scene dataset, where the overall mean Average Precision (mAP) for the target of interest, bridges, improved from 56.2% to 71.3%, achieving impressive results.**

I. Introduction

Detecting targets in large-scale optical remote sensing images has always been a challenging task, especially for tiny-sized targets. Generally, deep neural network-based object detectors [1]–[3] often struggle to capture robust feature information in tiny object detection (TOD) task, both semantically and spatially. Many researchers have proposed specific solutions for detecting tiny targets in ordinary and remote sensing images. For example, [4] uses a coarse-to-fine approach to improve the utilization of feature information within the receptive field by first screening high-potential proposals and then conducting fine classification in the second stage. [3], [5], [6]have proposed positive and negative sample learning strategies during the model training process to supply more high-quality positive samples that can be stably refined by the detector. [7] introduced the fountain fusion module to enhance the model's ability to fuse multi-scale features, promoting the extraction of spatial features from shallow feature layers. These methods have achieved good results in TOD tasks under different scenarios. However, their generalization performance in detecting small targets in large-scale remote sensing images has been unsatisfactory. Therefore, we deeply analyzed the issues within the network inference mechanisms for this detection scenario and proposed our solution, TOD-Net[1], to improve the efficiency and accuracy of the model in TOD scenarios.

In our application scenario, due to the large size of a single inference image, reaching 2000×2000 (to meet the inference speed requirements for large-scale remote sensing images), after a certain number of down-sampling layers in the feature extraction process, a single detection unit (DU) not only encompasses the target of interest but also cover a large area of the background, including the surrounding region of the target. The visualization results are shown in Fig 1.

We visualized the regions in the original input image corresponding to positive samples in a detection layer of size 256×256 (i.e., weight visualization). It can be seen that in the area where the target of interest—bridge appears, the effective receptive field (ERF) of this DU covers the target. However, smaller targets on the right side of the input image cannot generate effective positive samples (i.e., yellow points in Fig.1). This is due to the combined effect of the following two factors:

1) Foreground information of tiny remote sensing objects is easily submerged by the background during the down-sampling process of feature extraction..
2) Tiny objects are harder to highlight in the existing positive sample matching mechanisms, resulting in an inability to effectively and stably refine them.

Therefore, based on the coarse-to-fine feature extraction mechanism, we designed a vague search module (VGM) for filtering high-confidence proposal regions and combined it with centroid distance measurement (CDM) of polygonal areas to increase the number of positive detection units during model training. Based on the above, we formed the core algorithm mechanism of TOD-Net.

We tested TOD-Net on a self-collected remote sensing image dataset of an urban scene named GF2UBS (GF-2 urban bridge scene), where the images were captured by the GF-2 satellite and are single-band panchromatic images. Additionally, we validated the effectiveness and robustness of our algorithm on large public remote sensing datasets,

[1]This work was supported by the National Key Research and Development Program of China under Grant 2021YFB3900502.

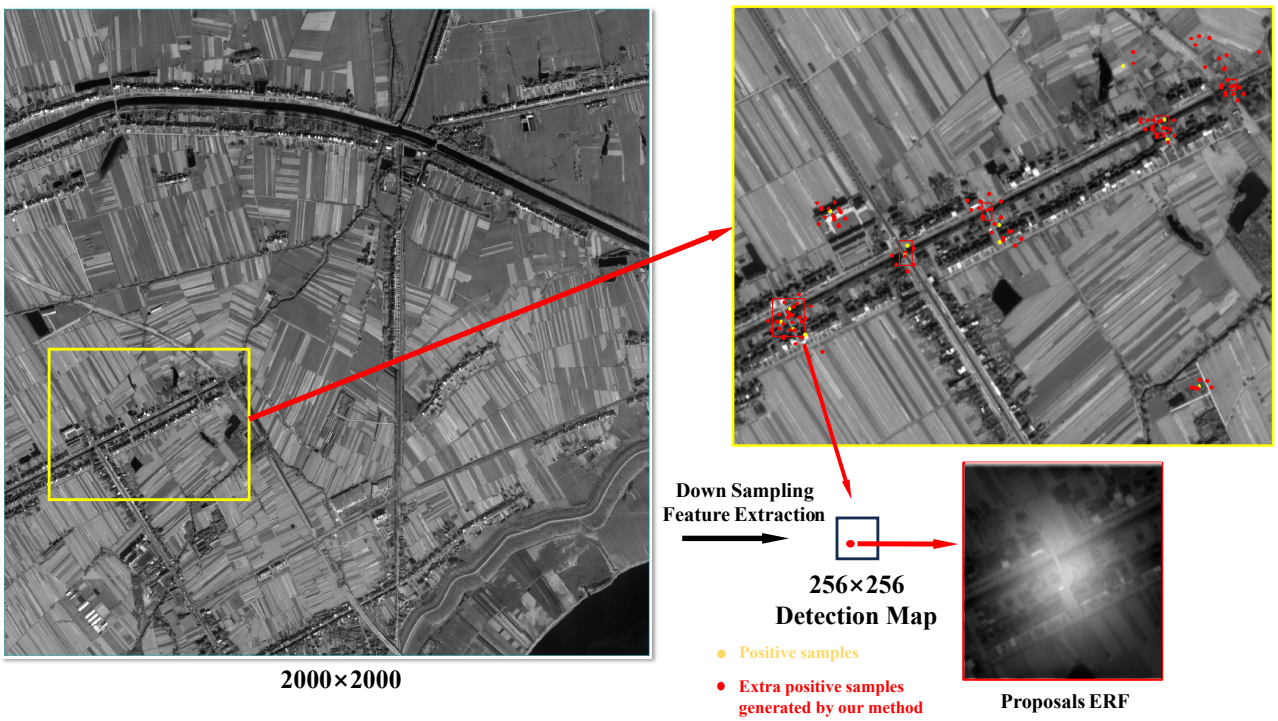

Fig. 1. Issues in detecting tiny objects of large-scale remote sensing imagery.

including the ship dataset HRSC2016 [8] and the multi-class remote sensing target dataset UCAS-AOD [9]. TOD-Net demonstrated superior performance in detecting small targets in large-scale remote sensing images, improving the overall detection accuracy (mAP) in the GF2UBS from 56.2% to an impressive 71.3%.

II. Methodology

A. Overall Pipeline of TOD-Net

TOD-Net directly uses large-sized remote sensing images (2000×2000) as network input during both training and testing phases. Consequently, under the same network configuration, the ERF of the DUs encompasses a larger spatial range of ground objects. The overall design process of TOD-Net is illustrated in Fig.2. After feature extraction through down-sampling by the deep neural network (DNN), multi-scale feature maps of varying sizes are obtained. TOD-Net adopts the classic ResNet [10] + FPN [11] combination as the backbone for feature extraction, followed by coarse-grained prediction box generation, resulting in the classification scores and regression results of the DUs. Then, two branches separately perform VGS and fine feature extraction, with the results of VGS serving as auxiliary information for the fine feature extraction step.

The multi-task overall loss in our experiments is defined as:

$$Loss_{Overall} = loss_{coarse} + loss_{fine}, \tag{1}$$

$loss_{coarse}$ and $loss_{fine}$ have the same format which is defined as:

$$loss_{coarse}, loss_{fine} = loss_{cls}(p, p^*) + loss_{reg}\left(\boldsymbol{t}, \boldsymbol{t}^*\right), \tag{2}$$

where the value p denotes the predicted classification score, p^* is the class label for all anchors ($p^* = 1$ for positive anchors and $p^* = 0$ for negative anchors). Symbol $\boldsymbol{t}$ is the position vector of predicted box offsets, and $\boldsymbol{t}^*$ denotes the GT box offsets. The position offsets vector $\boldsymbol{t}$ has the format (x, y, w, h, θ) where x and y denote the central coordinates, w the width, h the height, and θ the angle. For vector $\boldsymbol{t} = (t_x, t_y, t_w, t_h, t_\theta)$, we have:

$$\begin{aligned} t_x &= (x - x_a)/w_a, \quad t_y = (y - y_a)/h_a, \\ t_w &= \log(w/w_a), \quad t_h = \log(h/h_a), \\ t_\theta &= \tan(\theta - \theta_a), \end{aligned} \tag{3}$$

where x and x_a are the predicted box and anchor, respectively (likewise for y, w, h, θ). Focal Loss [12] is used in $loss_{cls}$ to reduce the contribution of easy-to-classify samples, and it alleviates the insufficient training caused by the class imbalance to a certain extent. The bounding box regression process (i.e., $loss_{reg}$) uses smooth L1 loss by default when there is no special statement.

B. VGS

For the obtained multi-scale feature maps $\mathbf{X}_i \in \mathbb{R}^{H\times W\times C}, i \in \{0, 1, 2, 3, 4\}$, i is the scale index in our TOD-Net. $\mathbf{X}_i(p)$ represents the feature vector corresponding to DU at position p in each feature map, which includes the confidence score of the predicted category generated in the coarse

detection stage and the regression parameters of the predicted box. MD_p is the positive and negative sample determination result at position p, $p \in \mathbb{P}\{0, 1, ..., W-1\} \times \{0, 1, ..., H-1\}$. For positive sample candidate regions, $MD_p = 1$, otherwise 0.

The core objective of VGS is to expand the search range within the coarse detection results where targets might appear, thereby maximizing the inclusion of valid targets within positive DUs that have larger receptive fields. This supports the feature extraction task in the subsequent fine-grained detection stage. As a result, searching expanding function $\tilde{\mathcal{F}}$ is defined as:

$$\begin{aligned} \tilde{\mathcal{F}} : MD_p = 1, \text{if } IoU(\mathbf{X}_i(p)) > 0.2, \\ \text{for } i \in [0, 5) \text{ and } p \in [p-1, p+1] \end{aligned} \tag{4}$$

where IoU is the traditional matching degree used in general object detectors. $\tilde{\mathcal{F}}$ is used to retrieve regions at any detection unit position that meet the criteria of having prediction confidence greater than 0.2 within all adjacent scales and surrounding 3×3 neighboring units. These proposals are then considered as positive samples, thereby generating potential candidate regions.

C. CDM

CDM is used to calculate the distance between the centroids of two polygonal regions as an evaluation metric. Its application is particularly suitable when the candidate region is large and the two regions have significant overlap characteristics. As shown in Fig. 4, the black dot $P_0(x_0, y_0)$ represents the centroid coordinates of the generated fuzzy candidate region, and the yellow dot $P_c(x_c, y_c)$ represents the centroid coordinates of the generated coupling region (CR). CR is the union of the predicted box (PB) generated during the coarse detection stage and the intersecting ground truth (GT). Therefore, the calculation method for Pc is:

$$x_c = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \textbf{and} \quad y_c = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{5}$$

After obtaining the centroid coordinates of CR, we can automatically determine whether the anchor box appearing in the refined feature extraction stage is a positive sample based on the inscribed circle radius of vague proposals, which is defined as:

$$CDM = Dis\left[P_0, P_c\right] <= \frac{\sqrt{2}}{2}\sqrt{\text{ratios}^{\,2}}, \tag{6}$$

where $ratios$ is the radius of the inscribed circle for vague proposals, Dis is the distance metric function, which can be measured using classical vector distance metrics such as Euclidean distance or Mahalanobis distance.

III. Experiments

A. Datasets

We first tested TOD-Net on a self-collected high-resolution optical remote sensing dataset, GF2UBS, which consists of GF-2 optical panchromatic images. The dataset contains 182

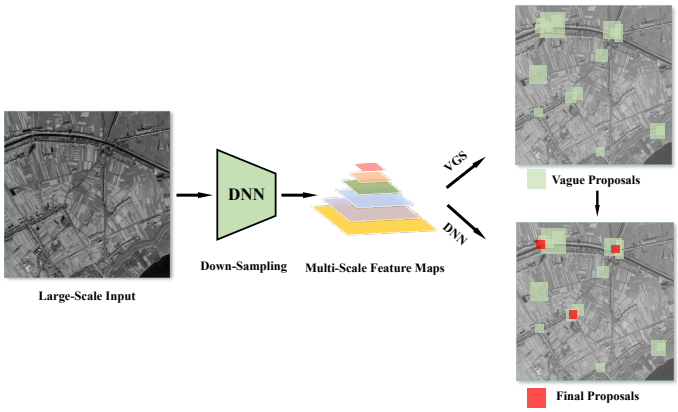

Fig. 2. Overall pipeline of TOD-Net. VGS denotes vague search strategy and we generate coarse detection results in the VGS process. CDM is used in the finer feature extraction process which cooperates with the output of VGS.

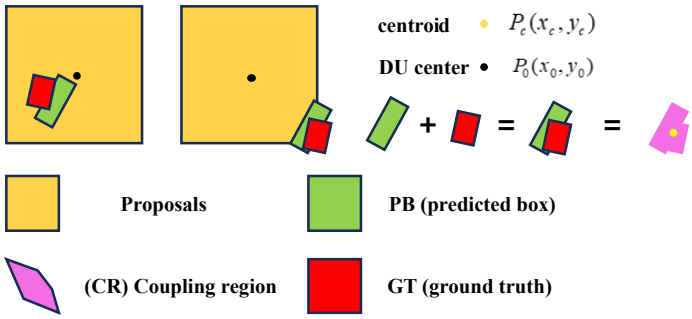

Fig. 3. Illustration of CDM.

panchromatic image slices, each sized 2000×2000 pixels. The dataset annotates 282 various types of bridges scattered throughout urban scenes, as shown in Fig. 4. Since most of the annotated bridges are small urban traffic bridges, their pixel sizes are relatively small, with a size distribution ranging from 7×10 to 121×36 pixels.

In addition to the remote sensing tiny object detection dataset in large-scale images, we also validated the generality of our method on two commonly used public remote sensing object detection datasets, which are introduced as follows:

1. HRSC2016 [8] is for ship detection that was collected from six harbors on Google Earth. It contains 1061 images (617 for training and 444 for testing). The image size ranges from 300 × 300 to 1500 × 900. We resize all images to 800 × 800.

2. UCAS-AOD [9] is an aerial image dataset for oriented aircraft and car detection which contains 1510 images including 1000 airplane images and 510 car images. We randomly divide it into the training set, validation set, and test set as 5:2:3.

B. Results on GF2UBS

Table I compares the detection accuracy results of TOD-Net with state-of-the-art object detection algorithms on the GF2UBS dataset. We used RetinaNet as the baseline for our model, with the baseline accuracy reaching only 56.2%. Under

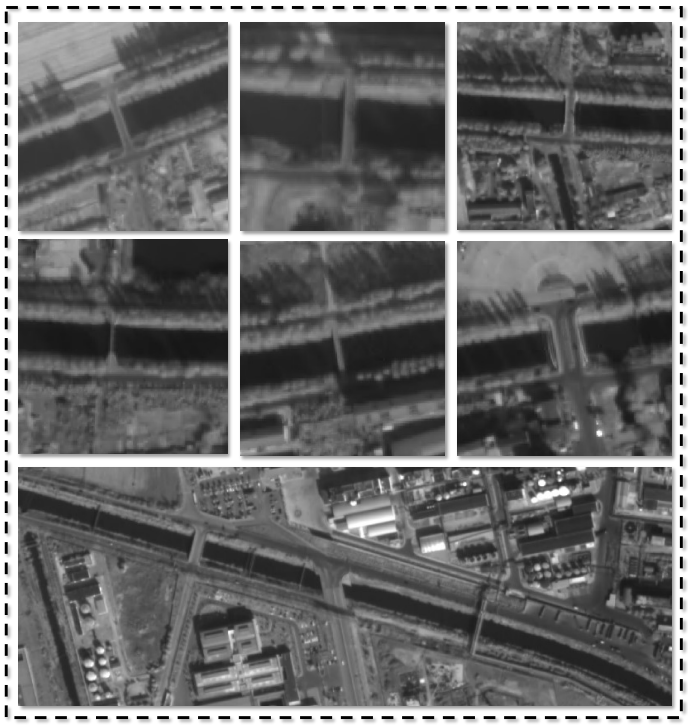

Fig. 4. Data examples of GF2UBS.

the same experimental configuration, TOD-Net increased the accuracy by 13.6%, which is quite impressive. This confirms the applicability and effectiveness of the proposed method in this paper.

When using more complex backbones for feature extraction, TOD-Net achieved a test accuracy of 71.3% on the GF2UBS, significantly outperforming other compared algorithms and achieving the fastest inference speed. Additionally, when using smaller image sizes as the benchmark for model training and testing, the mAP of TOD-Net further improved. This is because the model can capture more detailed spatial information despite reducing the network input size increases the overall inference time.

When focusing on the detection efficiency of tiny targets, the method proposed in this paper also achieved significantly better accuracy compared to other models. We evaluated the mAP-tiny of targets with pixel areas smaller than 30 across different algorithms. The performance of TOD-Net far surpassed that of the other methods, reaching a maximum of 40.1%. Although this result is not entirely satisfactory, it highlights the considerable difficulty in detecting small targets in the given scenes, suggesting there is still significant room for future improvement.

C. Comparison with State-of-Art Methods

To verify the generalizability of the method proposed in this paper, we also conducted validation on two additional public remote sensing datasets. On the HRSC2016 dataset, we compared our method with more state-of-the-art object detection algorithms, as shown in Table II and Table III. The

TABLE I
COMPARISONS WITH STATE-OF-THE-ART DETECTORS ON GF2UBS. BEST RESULTS ARE IN **BOLD**. RUNTIMES REPRESENTS THE TIME CONSUMED FOR MODEL INTERFERENCE OF TEST DATASET.

Methods	Backbone	Input Size	mAP-tiny	mAP-overall	runtimes (s)
S^2A-Net[13]	ResNet50	2000×2000	28.6	61.3	432.9
AOPG [14]	ResNet50	2000×2000	30.1	62.8	476.8
ReDet [6]	ResNet50	2000×2000	31.2	64.6	502.7
O-RCNN [15]	ResNet50	2000×2000	26.3	56.9	496.3
Baseline (RetinaNet)	ResNet50	2000×2000	24.6	56.2	465.4
TOD-Net (ours)	ResNet50	2000×2000	39.2	69.8	418.5
TOD-Net (ours)	ResNet50	800×800	44.3	76.5	538.5
TOD-Net (ours)	ResNet101	2000×2000	40.1	**71.3**	452.8

[1] mAp is calculated below VOC2012 rules.
[1] mAP-tiny refers to the detection precision of targets with a pixel area less than 30.

TABLE II
COMPARISONS WITH STATE-OF-THE-ART DETECTORS ON HRSC2016. NA DENOTES NUMBER OF ANCHORS. RUNTIMES REPRESENTS THE TIME CONSUMED FOR MODEL INTERFERENCE OF TEST DATASET.

Methods	Backbone	Input Size	NA	mAP (12)	runtimes (s)
*R^3Det-DCL [16]	ResNet101	800×800	-	96.41	134.6
*◇ DAL[1]	ResNet101	800×800	3	94.33	163.4
*◇ DRN[5]	ResNet101	800×800	-	95.65	154.1
*S^2A-Net[13]	ResNet50	800×800	1	95.01	128.6
*AOPG [14]	ResNet101	800×800	-	96.22	139.8
*ReDet [6]	-	800×800	-	97.63	167.1
*O-RCNN [15]	ResNet101	800×800	-	97.60	162.3
*CGC-Det[17]	ResNet101	800×800	-	97.86	168.3
TOD-Net (ours)	ResNet50	800×800	3	97.88	143.5
TOD-Net (ours)	ResNet101	800×800	3	**98.26**	152.8

[1] Best results are in **bold**.
[2] * represents the experimental runtimes are obtained from our local environment. ◇ indicates that the mAP is the result of re-realization.

results in these tables demonstrate that TOD-Net excels at enhancing the model's fine-grained feature extraction capabilities for targets in various scenes. This improvement effectively and consistently enhances the detector to extract robust features from target locations, thereby increasing the final accuracy.

On the HRSC2016 dataset, using ResNet101, we achieved a final mAP of 98.26%. On the UCAS-AOD dataset, we obtained the highest performances of 90.64% for car targets and 91.26% for airplane targets. These experimental results clearly demonstrate that TOD-Net not only performs excellently in large-scale remote sensing small target tasks but also adapts well to typical remote sensing target detection tasks in general application scenarios.

IV. CONCLUSIONS

In this paper, we propose a object detector for tiny objects in large-scale remote sensing images, named TOD-Net. TOD-Net employs a coarse-to-fine design and integrates a high-potential vague region screening strategy called VGS. VGS effectively expands the search range of detection units by searching for potential target areas across multiple scale levels and adjacent units. Additionally, we introduce a centroid distance metric to match small target positive samples, increasing the number of positive samples that can be refined during training.

TABLE III
DETECTION RESULTS ON UCAS-AOD DATASET.
BEST RESULTS ARE IN **BOLD**.

Methods	car	airplane	mAP
Baseline	0.8322	0.8643	0.8472
YOLOv3-O[18]	0.7463	0.8952	0.8208
Faster R-CNN-O [19]	0.8687	0.8986	0.8836
DAL[1]	0.8925	0.9049	0.8987
Oriented Reppoints [20]	0.8951	0.9070	0.9011
TOD-Net (ours)	**0.9064**	**0.9126**	**0.9106**

Experimental validation of TOD-Net on the large-scale scene remote sensing tiny target detection dataset GF2UBS shows an impressive 15.1% increase in the detection mAP for bridge targets, yielding remarkable results.

REFERENCES

[1] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," *arXiv preprint arXiv:2012.04150*, vol. 1, no. 2, p. 6, 2020.

[2] L. Ren, L. Gao, M. Wang, X. Sun, and J. Chanussot, "Hadgsm: A unified nonconvex framework for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[3] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic r-cnn: Towards high quality object detection via dynamic training," *arXiv preprint arXiv:2004.06002*, 2020.

[4] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 821–830.

[5] X. Pan, Y. Ren, K. Sheng, *et al.*, "Dynamic refinement network for oriented and densely packed object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 207–11 216.

[6] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2786–2795.

[7] T. Zhang, X. Sun, L. Zhuang, *et al.*, "Ffn: Fountain fusion net for arbitrary-oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[8] W. LB *et al.*, "A high resolution optical satellite image dataset for ship recognition and some new baselines," 2017.

[9] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *2015 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2015, pp. 3735–3739.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[13] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *arXiv preprint arXiv:2008.09397*, 2020.

[14] G. Cheng, J. Wang, K. Li, *et al.*, "Anchor-free oriented proposal generator for object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[15] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3520–3529.

[16] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection. arxiv 2020," *arXiv preprint arXiv:2011.09670*, 2021.

[17] Y. Wang, Z. Zhang, W. Xu, *et al.*, "Learning oriented object detection via naive geometric computing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2023. DOI: 10.1109/TNNLS.2023.3242323.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[19] G.-S. Xia, X. Bai, J. Ding, *et al.*, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[20] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1829–1838.