

Multi-scale Feature Enhanced Network for Remote Sensing Object Detection

Yanyan Zang¹, Xuewen Ding^{1,2,*}, Yuchen Jiang¹, Zewen Zi¹

¹ School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin, 300222, China

² Tianjin Yunzhitong Technology Co., Ltd., Tianjin, 300350, China

* Corresponding author: Xuewen Ding (Email: dingxw@tute.edu.cn)

Abstract: The task of remote sensing image target detection still faces many challenges: the target scale, orientation and spatial distribution are very different. In particular, small targets are difficult to identify in complex backgrounds, and the change of target geometry will affect feature extraction and positioning accuracy. In view of the above problems, this paper builds an improved detection network MSTD-Net based on YOLO26 m. Firstly, the global context Transformer module (GCTM) is introduced into the backbone network to capture the context information between long-distance spatial regions. Secondly, the dynamic deformable convolution module (DDCN) is used to adaptively adjust the convolution sampling position according to the geometric shape of the target, so that the network can better adapt to the targets with different shapes and orientations. In addition, this paper uses the dynamic Wise-IoU loss function (DWL) to adjust the contribution of training samples with different positioning quality in the bounding box regression process. In this paper, a large number of comparative experiments are carried out on DOTA15, UCAS-AOD and HRSC2016 datasets. The experimental results show that the mAP @ 50 index of MSTD-Net on the three datasets reaches 7575 %, 9730 % and 7403 %, respectively. Compared with the baseline model YOLO26m, MSTD-Net increased mAP @ 50 by 100 %, 026 % and 792 % on DOTA15, UCAS-AOD and HRSC2016 datasets, respectively. The ablation experiment further verifies that the three improved modules all have gains in detection performance, but the improvement is different on different data sets.

Keywords: Remote Sensing Images; Target Detection; YOLO; Transformer; Deformable Convolution.

1. Introduction

Aerial images usually cover a large geographical area, and a single target may only account for a small part of the image, or it may be densely distributed with other targets. The difference of target size, direction and background conditions will increase the difficulty of target recognition and positioning. Existing remote sensing target detection research and aerial image benchmark datasets have discussed these issues [1,2].

Deep learning provides different technical approaches for target detection. Faster R-CNN [3] uses a two-stage detection process to first generate candidate regions, then classify the targets and refine the bounding box. In contrast, YOLO [4] performs detection within a unified prediction framework, providing another trade-off between detection accuracy and computational efficiency. However, it is still challenging to accurately detect small targets in complex aerial background, which puts forward higher requirements for feature extraction ability.

Target detection in remote sensing images is mainly affected by two aspects. The first is the target scale change. Small targets contain limited visual information, while larger targets may need to use features that cover a wider spatial range. The second is the geometric change of the target. Aircraft, ships and vehicles may appear in images in different directions, and convolutional layers with fixed sampling modes are difficult to represent the structural features of these targets equally well. Therefore, effective feature extraction needs to consider both the context information around the target and the local geometric features of the target.

In response to the above problems, existing research has explored from different perspectives. Feature Pyramid

Network (FPN) achieves feature representation of different scale targets by fusing high-level semantic information and low-level spatial features [5]. Vision Transformer (ViT) [6] and Swin Transformer [7] use the attention mechanism to establish the relationship between different regions of the image. The deformable convolutional network [8,9] adopts another way to adjust the sampling mode according to the input features by learning the offset of the convolution sampling position. These methods improve the ability of feature representation in different ways. However, for aerial target detection, how to retain the geometric details of the target itself while using context-related information is still a matter of concern.

Bounding box regression is also an important factor affecting the performance of target detection. Generalized IoU (GIoU) [10] extends the traditional IoU metric by considering the minimum circumscribed region of two bounding boxes. Distance-IoU (DIoU) [11] further introduces the center distance between the prediction box and the real box. Wise-IoU (WIoU) [12] proposed a dynamic focusing mechanism to adjust the contribution of different training samples in the bounding box regression optimization process. These studies provide a reference for the design of remote sensing target location loss function that can consider the difference of sample quality.

Based on the previous analysis, this paper proposes a remote sensing target detection network MSTD-Net based on YOLO26 m. The network integrates context feature modeling, adaptive spatial sampling and bounding box regression optimization into the same detection framework. The main work of this paper is as follows:

(1) The Global Context Transformer Module (GCTM) is introduced into the backbone network to establish the

relationship between different spatial regions and extract context features.

(2) The Dynamic Deformable Convolution Module (DDCN) is introduced to adjust the convolution sampling position according to the geometric features of the target to adapt to different shapes and directions of the target.

(3) The dynamic Wise-IoU loss (DWL) function is used to adjust the contribution of different training samples to the bounding box regression according to the positioning quality. The proposed modules were evaluated by comparative experiments and ablation experiments on DOTA15, UCAS-AOD and HRSC2016 datasets.

2. MSTD-Net

2.1. Global Context Transformer Module

Large aerial scenes often have cluttered backgrounds, and targets may be scattered across the area. Standard convolution mainly gathers local features, which can limit information flow between distant image regions. To strengthen global

feature representation, this study brings in the Global Context Transformer Module (GCTM). Figure 1 shows the structure of GCTM.

For an input tensor, the channel representation is first transformed by a 1×1 convolution. The resulting feature is then split into several regions and fed into Transformer blocks. Within each processed region, self-attention calculates the relationships between feature positions, as given in Equation (1). In Equation (1), Q , K , and V stand for the query, key, and value matrices, respectively. Residual fusion produces the output feature, as shown in Formula (2).

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \quad (1)$$

$$F_{\text{out}} = \text{Conv}(F_{\text{att}} + X) \quad (2)$$

GCTM allows information to be exchanged across spatial regions. It also brings contextual cues into the feature representation used for detection in cluttered aerial scenes.

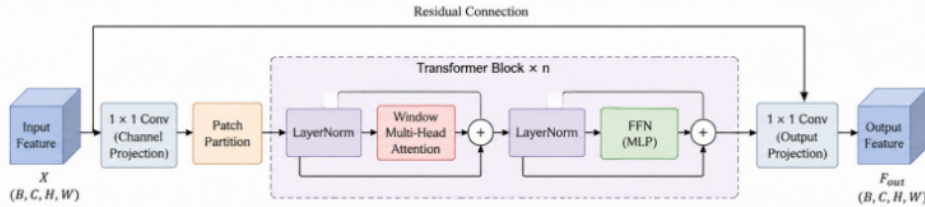


Figure 1. Structure of GCTM

2.2. Dynamic Deformable Convolution Module

Targets in aerial imagery often vary a lot in size, shape, and orientation. Vehicles, ships, and aircraft, for instance, can face different directions and have irregular structures. Standard convolution relies on a predefined sampling grid, so its sampling pattern responds poorly when object geometry changes. To improve how feature sampling adapts to object geometry, MSTD-Net adds a Dynamic Deformable Convolution Module (DDCN). The architecture of DDCN is shown in Figure 2. Formula (3) gives the standard convolution operation.

$$y(p) = \sum_{k \in R} w(k)x(p+k) \quad (3)$$

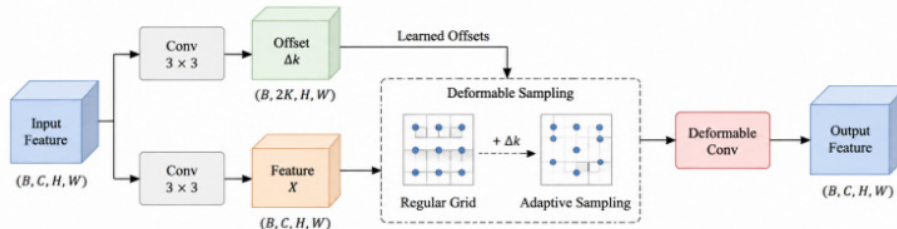


Figure 2. Structure of DDCN

2.3. Dynamic Wise-IoU Loss

Bounding box regression decides how predicted regions are shifted toward the ground-truth annotations they match. Aerial objects differ in size and shape, so localization errors may vary a lot from one predicted bounding box to another. Conventional IoU-based objectives mainly quantify the overlap between predicted and reference boxes, yet they often treat different training samples equally, which can limit optimization for complex samples. Figure 3 shows how DWL

is calculated.

To account for differences in regression quality, MSTD-Net adopts a Dynamic Wise-IoU Loss (DWL). Related sample-weighting strategies have been investigated in Wise-IoU [12]. The proposed loss introduces a dynamic quality evaluation mechanism to measure the regression quality of different samples. For a predicted bounding box B_p and its corresponding ground-truth box B_g , their overlap is calculated using Equation (5).

$$\text{IoU} = \frac{|\mathbf{B}_p \cap \mathbf{B}_g|}{|\mathbf{B}_p \cup \mathbf{B}_g|} \quad (5)$$

The overlap measurement is subsequently supplemented with a geometric constraint. The matching quality score is obtained according to Formula (6).

$$S_W = \text{IoU} - \left(\frac{\rho^2}{c^2} + \alpha v \right) \quad (6)$$

where, ρ denotes the distance between the two box centers, c represents the diagonal length of their minimum enclosing box, and v is the aspect-ratio consistency term.

Based on the matching quality S_W , a sample-dependent focusing factor determines the relative contribution of each regression instance during training. Its calculation is given in Equation (7).

$$r_i = \text{Clamp} \left(\frac{q_i}{\bar{q} + \epsilon}, 0.5, 1.5 \right) \quad (7)$$

In Equation (7), q_i denotes the current sample's quality score, $\bar{q}$ is the mean quality score within the batch, and ϵ prevents the denominator from becoming zero. The dynamic weighting strategy is introduced into the regression loss function, as expressed in Formula (8).

$$L_{\text{DWL}} = r_i(1 - S_W) \quad (8)$$

The dynamic weighting factor adjusts the contribution of each sample to the bounding box regression loss. The effect of this mechanism on detection performance is examined through the ablation experiments presented in Section 3.5.

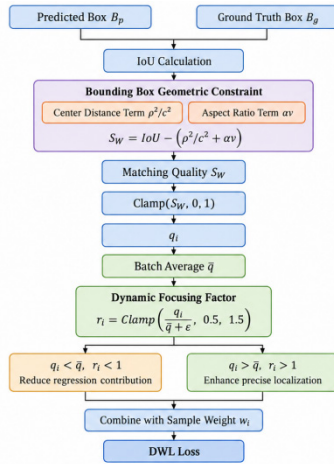


Figure 3. Flowchart of DWL.

3. Experiments and Analysis

3.1. Experimental Environment

The experiments were implemented using Ultralytics and PyTorch. The software environment consisted of Windows, Python 3.10.18, Ultralytics 8.4.31, PyTorch 2.10.0, and CUDA 12.6. Training was performed on an NVIDIA GeForce RTX 4080 SUPER GPU with 16 GB of memory.

The input images were set to 640×640 pixels, and the model was trained for 100 epochs with a batch size of 16. SGD was used with an initial learning rate of 0.01. The training pipeline included horizontal and vertical flipping, random rotation, Mosaic, MixUp, and Copy-Paste.

3.2. Datasets

In this paper, three remote sensing target detection data sets of DOTA15, HRSC2016 and UCAS-AOD are selected for

experiments.

The DOTA dataset contains aerial targets with different scales, directions, and distribution densities [2]. Compared with DOTA-v10, DOTA15 adds minimal target labeling and container crane categories, including 16 target categories and 403,318 labeling instances. Due to the large difference in target size and the complexity of the scene, the data set can be used to test the multi-scale detection performance of the model.

HRSC2016 is mainly used for ship detection in high-resolution remote sensing images [13]. The ships in the data set are different in size, aspect ratio and orientation, and the sea surface background of some images is also more complex, which can be used to investigate the detection effect of the model on different forms of ships.

UCAS-AOD includes two categories : aircraft and vehicle. [14] The size, direction and background of the target are different. Compared with DOTA1.5, the data set has fewer target categories and can be used as another set of experimental data to compare the detection performance of the model in different aerial scenes.

Before training, the image is uniformly adjusted to 640×640 pixels, and the annotation is converted to YOLO format, including target category and normalized bounding box coordinates. The data augmentation method described in Section 3.1 is used in the training process.

3.3. Evaluation Metrics

This study assessed detection performance with mean Average Precision at an Intersection over Union (IoU) threshold of 0.5 (mAP@50). Average Precision was computed for each category, then averaged across categories in the evaluated dataset; the comparative experiments and the ablation studies both adopted this same metric.

3.4. Comparative Experimental Results

Table 1 compares MSTD-Net with Faster R-CNN, RetinaNet, and YOLOv8 on DOTA1.5, UCAS-AOD, and HRSC2016 using mAP@50.

Table 1. Performance Comparison with Mainstream Models

Method	Backbone	DOTA1.5	UCAS-AOD	HRSC2016
		mAP@50	mAP@50	mAP@50
Faster R-CNN	ResNet-50	66.94	89.02	68.87
RetinaNet	ResNet-50	62.73	85.43	70.76
YOLOv8	YOLOv8	68.65	92.60	73.22
MSTD-Net	YOLO26m	75.75	97.30	74.03

According to Table 1, MSTD-Net obtains mAP@50 scores of 75.75%, 97.30%, and 74.03% on DOTA1.5, UCAS-AOD, and HRSC2016, respectively. On DOTA1.5, these results exceed Faster R-CNN (66.94%), RetinaNet (62.73%), and YOLOv8 (68.65%) by 8.81, 13.02, and 7.10 percentage points. For UCAS-AOD, the corresponding results of the three reference methods are 89.02%, 85.43%, and 92.60%. On HRSC2016, MSTD-Net exceeds YOLOv8 (73.22%) by 0.81 percentage points.

The improvements are mainly attributed to the three proposed components. GCTM enhances global contextual information extraction, DDCN improves the adaptability to geometric variations of remote sensing objects, and DWL optimizes the bounding box regression process. These improvements enable MSTD-Net to achieve more accurate

feature representation and localization performance compared with existing detection methods.

3.5. Ablation Study

Table 2 presents the DOTA1.5 ablation experiments based on YOLO26m. GCTM, DDCN, and DWL were tested individually, followed by the GCTM+DDCN combination and the complete MSTD-Net.

Table 2. Ablation study of different components on the DOTA1.5 dataset.

Method	mAP@50(%)	Δ mAP@50(%)
Baseline	74.75	-
+GCTM	75.12	+0.37
+DDCN	75.06	+0.31
+DWL	74.93	+0.18
+GCTM+DDCN	75.41	+0.66
MSTD-Net	75.75	+1.00

The baseline achieves 74.75% mAP@50 on DOTA1.5. With GCTM, DDCN, and DWL added separately, the scores reach 75.12%, 75.06%, and 74.93%, giving gains of 0.37, 0.31, and 0.18 percentage points. The GCTM+DDCN configuration reaches 75.41% (+0.66 points), while the complete network reaches 75.75% (+1.00 point). Thus, the full configuration scores 0.34 points higher than the two-module configuration under this evaluation setting.

The same ablation configurations were evaluated on UCAS-AOD; their mAP@50 scores are given in Table 3.

Table 3. Ablation study of different components on the UCAS-AOD dataset.

Method	mAP@50(%)	Δ mAP@50(%)
Baseline	97.04	-
+GCTM	97.15	+0.11
+DDCN	97.12	+0.08
+DWL	97.09	+0.05
+GCTM+DDCN	97.22	+0.18
MSTD-Net	97.30	+0.26

On UCAS-AOD, the baseline score is 97.04% mAP@50. Adding GCTM, DDCN, or DWL separately gives 97.15%, 97.12%, or 97.09%, corresponding to increases of 0.11, 0.08, or 0.05 percentage points. The GCTM+DDCN variant reaches 97.22% (+0.18 points), and MSTD-Net reaches 97.30% (+0.26 points). The absolute gains are smaller than those recorded on HRSC2016, but the reported results do not establish the reason for this difference.

Table 4 gives the corresponding ablation results on HRSC2016.

Table 4. Ablation study of different components on the HRSC2016 dataset.

Method	mAP@50(%)	Δ mAP@50(%)
Baseline	66.11	-
+GCTM	69.85	+3.74
+DDCN	69.23	+3.12
+DWL	67.64	+1.53
+GCTM+DDCN	72.17	+6.06
MSTD-Net	74.03	+7.92

The HRSC2016 baseline records 66.11% mAP@50. The

single-module results are 69.85% for GCTM (+3.74 points), 69.23% for DDCN (+3.12 points), and 67.64% for DWL (+1.53 points). The combination of GCTM and DDCN lifts the score to 72.17%, which is 6.06 points higher than before. When all components are included, the full model achieves 74.03%; this represents an improvement of 7.92 percentage points over the baseline, and it also exceeds the GCTM+DDCN variant by 1.86 points. Among the three datasets, HRSC2016 exhibits the largest absolute gain relative to its corresponding baseline.

3.6. Visualization Comparison of Detection Results

Figure 4 shows the comparison of the detection results of YOLO26 m baseline model and MSTD-Net on three remote sensing images. In each group of images, the left side is the detection result of YOLO26 m, and the right side is the corresponding detection result of MSTD-Net.

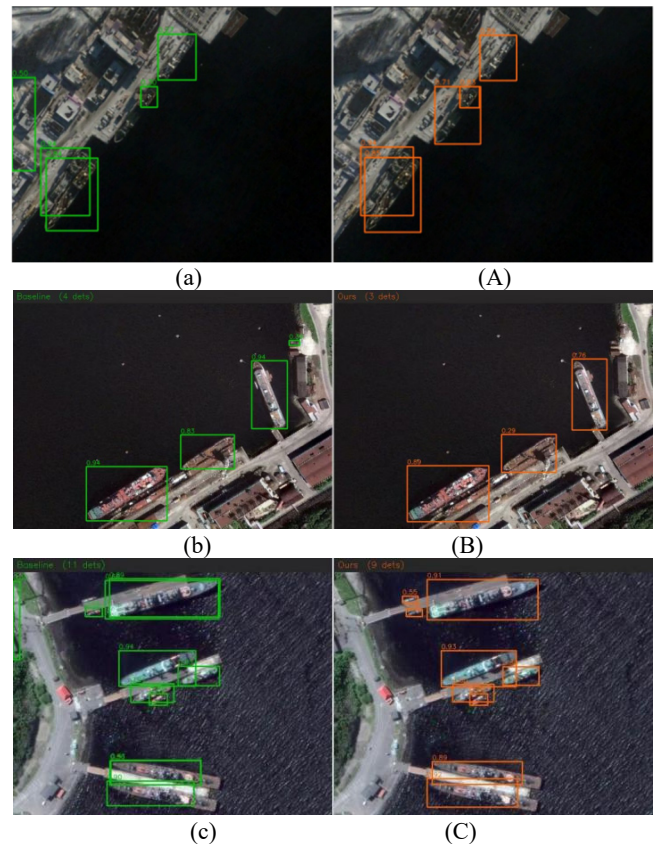


Figure 4. The detection results of YOLO26 m baseline model (a-c) and MSTD-Net (A-C).

As shown in Figure 4, the YOLO26 m baseline model and MSTD-Net show different detection results on selected remote sensing images. In some scenarios, the baseline model has target missed detection, while MSTD-Net can detect more targets. At the same time, in the figure, the prediction box generated by MSTD-Net has a better matching effect with the target area, especially when the target scale changes greatly and the background is more complex. These qualitative results are consistent with the performance improvement trend reflected in quantitative experiments

4. Conclusion

This work develops MSTD-Net, an improved aerial object detection network built on YOLO26m, to address object scale

variation, complex backgrounds, and weak feature representation in remote sensing images. The global context Transformer module GCTM is introduced into the backbone of the network to strengthen the information interaction between different spatial regions to improve the feature extraction under complex background. Considering the difference in size and shape of remote sensing targets, a dynamic deformable convolution module DDCN is used to adjust the convolution sampling position according to the geometric characteristics of the target. In addition, the dynamic Wise-IoU loss function DWL is introduced in the bounding box regression, and its weight in the loss calculation is adjusted according to the positioning quality of the sample.

The experiment was carried out on three datasets of DOTA15, UCAS-AOD and HRSC2016. The overall detection performance of MSTD-Net is better than that of the baseline network. Ablation experiments show that GCTM, DDCN and DWL modules can each bring mAP @ 50 improvement. Compared with the YOLO26 m baseline, the mAP @ 50 of MSTD-Net on the three datasets increased by 100, 026 and 792 percentage points, respectively. These results have certain reference value for the subsequent optimization of remote sensing image target detection model.

References

- [1] Cheng, G., & Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117, 11–28. <https://doi.org/10.1016/j.isprsjprs.2016.03.004>.
- [2] Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2018). DOTA: A large-scale dataset for object detection in aerial images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (pp. 3974–3983). <https://doi.org/10.1109/CVPR.2018.00423>.
- [3] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 91–99. <https://doi.org/10.48550/arXiv.1506.01497>.
- [4] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (pp. 779–788). <https://doi.org/10.1109/CVPR.2016.91>.
- [5] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (pp. 2117–2125). <https://doi.org/10.1109/CVPR.2017.106>.
- [6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2010.11929>.
- [7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (pp. 10012–10022). <https://doi.org/10.1109/ICCV48922.2021.00986>.
- [8] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (pp. 764–773). <https://doi.org/10.1109/ICCV.2017.89>.
- [9] Zhu, X., Hu, H., Lin, S., & Dai, J. (2019). Deformable ConvNets v2: More deformable, better results. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (pp. 9308–9316). <https://doi.org/10.1109/CVPR.2019.00953>.
- [10] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (pp. 658–666). <https://doi.org/10.1109/CVPR.2019.00074>.
- [11] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding box regression. In *Proc. AAAI Conf. Artif. Intell.*, 34(7), 12993–13000. <https://doi.org/10.1609/aaai.v34i07.6999>.
- [12] Tong, Z., Chen, Y., Xu, Z., & Yu, R. (2023). Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv preprint arXiv:2301.10051*. <https://doi.org/10.48550/arXiv.2301.10051>.
- [13] Liu, Z., Wang, H., Weng, L., & Yang, Y. (2016). Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds. *IEEE Geoscience and Remote Sensing Letters*, 13(8), 1074–1078. <https://doi.org/10.1109/LGRS.2016.2573468>.
- [14] Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., & Jiao, J. (2015). Orientation Robust Object Detection in Aerial Images Using Deep Convolutional Neural Network. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 3735–3739). <https://doi.org/10.1109/ICIP.2015.7351491>.