Local Attention Networks for Occluded Airplane Detection in Remote Sensing Images

Min Zhou#, Zhengxia Zou#, and Zhenwei Shi*, *Member IEEE*, Wen-Jun Zeng, *Member IEEE*, and Jie Gui, *Senior Member, IEEE*

***Abstract*—Despite the great progress of deep learning and target detection in recent years, the accurate detection of the occluded targets in remote sensing images still remains a challenge. In this letter, we propose a new detection method called local attention networks to improve the detection of occluded airplanes. Following the idea of "divide and conquer", the proposed method is designed by first dividing an airplane target into four visual parts: head, left/right wings, body and tail, and then considering the detection as the prediction of the individual key points in each of the visual part. We further introduce an additional attention branch in the standard detection pipeline to enhance the features and make the model focus on individual parts of a target even it is only partially visible in the image. Detection results and ablation studies on three remote sensing target detection datasets (including two publicly available ones) demonstrate the effectiveness of our method, especially for occluded airplane targets. In addition, our method outperforms the other state of the art detection methods on these datasets, such as Faster RCNN, SSD, and RefineDet.**

***Index Terms*—Airplane Detection, Target Occlusion, Remote Sensing Images, Attention Mechanism**

I. Introduction

The automatic detection of airplane targets in remote sensing images has long been playing an important role in remote sensing image analysis due to both of its military and civil use. In recent years, the fast development of deep learning technology [1] has greatly promoted the progress of airplane target detection [5–10] which has also received increasing attention.

The existing airplane detection methods can be roughly divided into two groups, traditional airplane detection methods and deep learning based methods, where the former are designed with sophisticated handcrafted features, visual saliency, and detection rules [2–4], while the latter are designed on basis of the recent popular deep learning detection architectures [5–8] (e.g. RCNN [11], Faster RCNN [12], SSD [20], etc). Despite the recent advances, the detection of the occluded targets in remote sensing images has long been difficult and still remains a challenge. As reported by C. Stubenrauch et al [25], on average, more than 50% of the earth's surface is covered by clouds every day, a target in optical remote sensing images has a large probability being covered by clouds. Besides, it is also likely for an airplane target being partially occluded by the roof of a hangar. In these conditions, it becomes harder for the classical detection frameworks to obtain satisfied detection results.

The work was supported by the National Key R&D Program of China under the Grant 2017YFC1405605, the National Natural Science Foundation of China under the Grant 61671037, the Beijing Natural Science Foundation under the Grant 4192034 and the National Defense Science and Technology Innovation Special Zone Project. (*Corresponding author: Zhenwei Shi*).

# These authors contributed equally to this work and are co-first authors.

Min Zhou and Zhenwei Shi (Corresponding author, e-mail: shizhenwei@buaa.edu.cn) are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.

Zhengxia Zou, Wen-Jun Zeng and Jie Gui are with the Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

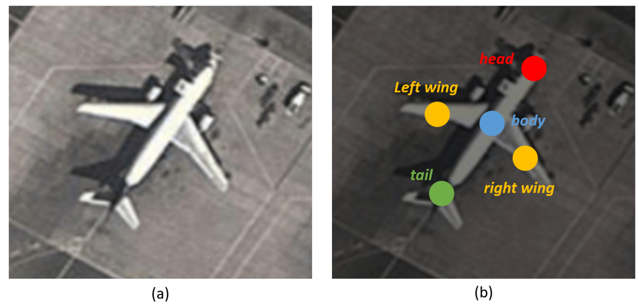

Fig. 1. (Better viewed in color) (a) An airplane image. (b) The different components (head, wings, body, and tail) and corresponding key points.

In the last two years, some researchers have proposed some methods for occluded airplane detection [13, 14] based on handcrafted features and traditional machine learning techniques. However, few people have done relevant research on deep learning based detection methods. Considering the above problems, in this letter, we propose a new deep learning based method to improve the detection of occluded airplane targets in remote sensing images. The proposed method is designed based on the fact that although a target is partially occluded, some other parts are still visible and can be identified as important visual cues for detection. This idea, which is also known as "detection by parts", has a long history in target detection, and has played an important role in traditional handcrafted detection methods, such as the deformable part based model [15]. We embrace this idea by taking advantage of the deep convolutional networks and make improved modern deep learning based detectors for occluded target detection problems. To enhance the features and make the model concentrate on visible parts of occluded targets, the attention mechanism is further integrated into our method. The attention mechanism was originally introduced in machine

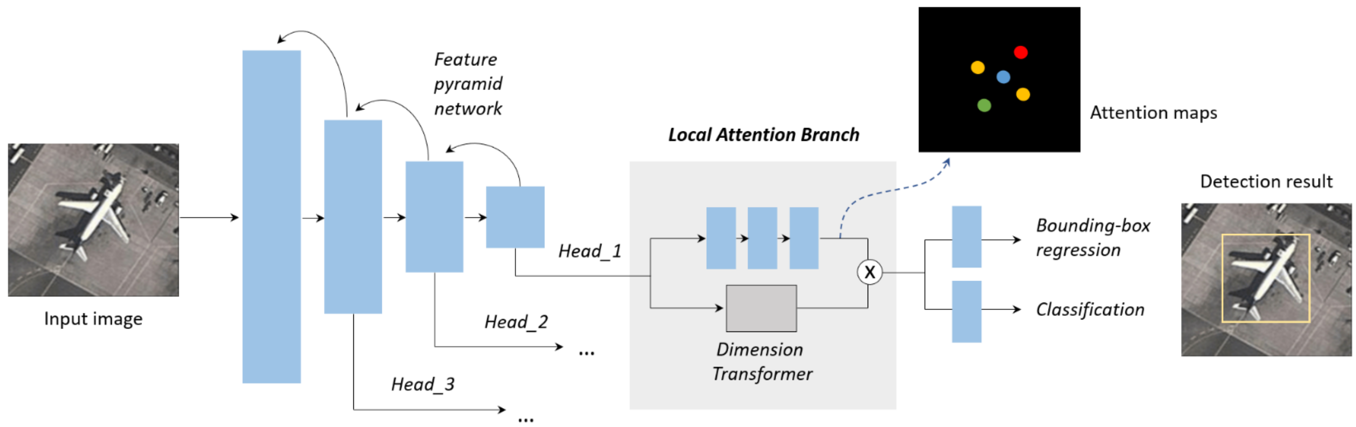

Fig. 2. An overview of the proposed local attention network. Our network consists of a standard feature pyramid network [16] and a local attention branch. The attention branch is used to enhance the features and make model focus on different parts of a target so that to improve occluded target detection. The dimension transformer aims to transform the channel dimension of the feature maps by performing a set of 1x1 convolutions on them so that to ensure the number of feature map channels can be divided exactly by the number of key points.

translation to improve the performance of an Encoder-Decoder RNN model by taking into account the input from several time steps to make one prediction [26]. In a CNN-based model, the introduction of attention mechanism is helpful for investigating the spatial correlations of different feature locations and now has been widely used in many computer vision tasks. In our method, we further introduce an additional attention branch in the traditional detection pipeline so that to highlight the features of the individual parts of targets and suppress those of the backgrounds. We called the detection network equipped with the attention branch "Local Attention Networks (LAN)".

The rest of this paper is organized as follows. We give a detailed introduction to our method in Section II. The experimental results are given in Section III, and the conclusions are drawn in Section IV.

II. Methodology

In this section, we will introduce the proposed local attention networks with details.

A. Overview

Fig. 2 shows an overview of our method. Our method is designed based on a well-known detection system, "Retina Net" [17] which is a State-Of-The-Art (SOTA) one-stage detection architecture in general object detection tasks.

Our network consists of a standard feature pyramid network [16] for feature extraction, and an additional local attention branch for integrating attention mechanism into the feature representation. There are three individual detection heads in our method. Each of them consists of three brunches: a local attention brunch, a detection brunch, and a classification branch. The input of each head is the feature maps extracted by the backbone of the networks while the outputs are a set of bounding boxes, which are further processed by non-maximum suppression to get the final detection results.

We consider the detection as the prediction of the individual key points in each of the visual part. According to an airplane's visual appearance, we split it into five parts, as shown in Fig. 1: a head, a left wing, a right wing, a body, and a tail. Then we set key points on the center of all parts. There are four types of key points shown with different colors in Fig. 1. The red, orange, blue and green points represent heads, wings, bodies, and tails respectively. As the left and right wings are mirror symmetry, we simply set one kind of key points for these two types of parts. On this basis, the attention branch is designed to enhance the features learned by the standard feature pyramid network and help to recall occluded target. More specifically, our net detects parts of airplanes, transform detection results into attention maps and enhance the features by making pixel-wise production between the attention maps and the features. This makes the network focus on target regions and meanwhile utilize detection results of airplane parts to help detection.

B. Local Attention Branch

As shown in Fig. 3, the local attention branch predicts a four-channel attention map for four different key points. When we perform detection, each local attention branch produces a 4-channel attention maps for a single input image, one channel per visual part. The features are then transformed by multiplying the attention maps to make it focus on different parts of the airplane. In addition to the loss of standard detection pipeline, the predicted tensor is compared with the label to compute loss and back propagate gradients.

Specifically, we evenly divide the feature maps into four groups along their channel dimensions and each group of feature maps are multiplied with a corresponding group of key points i ($i \in \{\text{head, wing, body, tail}\}$). Then the feature maps are concatenated together as the final enhanced feature maps. We do not set a key point as a single pixel, but a circle around the geometrical center. The pixel values of the key point maps is associated with the "distance" between the pixel to the center of a visual part, where the large value is, the closer they will be. We set the value of each element in the

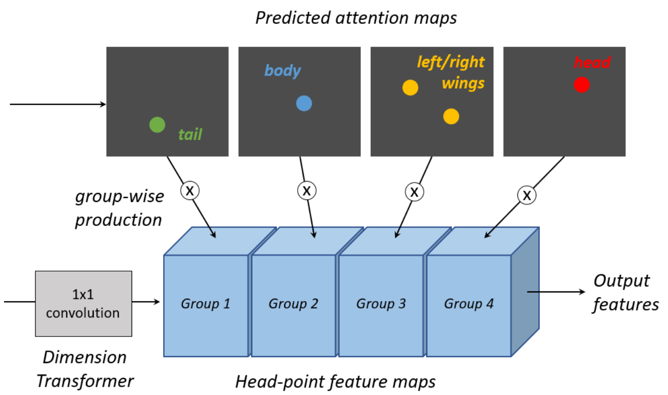

Fig. 3. The local attention branch predicts a four-channel attention map for four different key points, and then highlight target parts by making group-wise production between this attention map and transformed features.

key point label map as follows:

$$\hat{m}^{(s)} = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-r_s^2/2\sigma^2} & \text{if} \quad r_s^2 < (3\sigma)^2 \\ 0 & \text{else,} \end{cases} \tag{1}$$

where $r_s^2 = (x^{(s)} - x_c)^2 + (y^{(s)} - y_c)^2$ represent the distance of current pixel to the center coordinate (x_c, y_c) of a certain key point. σ is set to 1/20 size of the ground truth bounding box it belongs to. Besides, each attention map corresponds to only one type of key point to distinguish different parts. For example, in the tail attention map, the pixel values of the background and other type key points are set to zero.

C. Loss Function Design

In our method, each detection head consists a number of predefined anchor boxes and performs detection independently and has the same configurations except for the anchor size. The loss for each anchor further consists of three parts, 1) the loss for attention map prediction, 2) the loss for classification and 3) the loss for bounding box regression.

- Attention map prediction loss

The learning and prediction of the attention maps can be formulated as a pixel-wise regression problem. We use pixel-wise mean square error as the loss function. Suppose m is the predicted output of a certain pixel from one of the head points of the net, and $\hat{m}$ is its prediction label, then the loss function for this pixels can be calculated as

$$L_{point\_reg}(m, \hat{m}) = (m - \hat{m})^2, \tag{2}$$

- Bounding box regression loss

We use the smooth L_1 loss, which has been used in Faster-RCNN [18] as our bounding box regression loss function,

$$L_{bb\_reg}(\boldsymbol{t}, \hat{\boldsymbol{t}}) = \sum_{k \in \{x,y,w,h\}} SL_1(t_k - \hat{t}_k), \tag{3}$$

where $\boldsymbol{t}^{(i)}$ and $\hat{\boldsymbol{t}}^{(i)}$ represent the normalized coordinates [18] of the predicted and groundtruth bounding boxes (the top-left points, bounding box height and width). The smooth L1 function is used as the regression loss:

$$SL_1(x) = \begin{cases} 0.5x^2 & \text{if} \quad |x| < 1, \\ |x| - 0.5 & \text{else.} \end{cases} \tag{4}$$

- Anchor category classification loss

For detecting airplanes, the discrimination of targets and backgrounds can be considered as a binary classification problem. We use the focal loss [17] as the loss function. Suppose the classification output $\boldsymbol{p} = [p_0; p_1]$ represents the predicted probabilities of the background and target. Suppose $\boldsymbol{y} = [y_0; y_1]$ represents the one-hot label vector. Then the probability of misclassification can be written as:

$$p_{miss} = 1 - (y_0 p_0 + y_1 p_1). \tag{5}$$

For each anchor, its focal loss can be calculated by reshaping the standard cross-entropy loss so that to make it focus more on hard examples:

$$L_{cls}(\boldsymbol{p}, \boldsymbol{y}) = -p_{miss}^{\gamma}(\alpha_0 y_0 \log p_0 + \alpha_1 y_1 \log p_1), \tag{6}$$

where γ is a positive factor that controls the shape of the loss function.

- Final multi-task loss function

The total loss of our networks can be written as the linear combination of the losses of the above three tasks:

$$\begin{aligned} L_{total} = {} & \beta_1 \sum_s L_{point\_reg}(m^{(s)}, \hat{m}^{(s)}) \\ & + \beta_2 \sum_i \mathbb{1}\{a^{(i)} = 1\} L_{bb\_reg}(\boldsymbol{t}^{(i)}, \hat{\boldsymbol{t}}^{(i)}) \\ & + \beta_3 \sum_i L_{cls}(\boldsymbol{p}^{(i)}, \boldsymbol{y}^{(i)}), \end{aligned} \tag{7}$$

where i represents the id of an anchor, s represents all possible pixel locations of the feature map. $\mathbb{1}\{a = 1\}$ is an indicator function, meaning that only when the anchor is positive, its output value is 1 and otherwise 0. β_1, β_2 and β_3 are parameters to balance the weights between different tasks. The three tasks thus can be trained at the same time in an end-to-end fashion.

III. Experimental Results and Analysis

A. Dataset and Evaluation Metric

We use three remote sensing target detection datasets to conduct our experiments. One of them is collected by ourselves and the other two are publicly available datasets: NWPU10 [23] and occluded airplane datasets [13]. The statistics of these three datasets are listed in Table I. Particularly, in the dataset [13], there are 184 annotated airplane targets where 96 of them are occluded or truncated ones. The authors only release its testing set. The training set is not available. In our experiments, we randomly select 70% images from our own data set for training. The rest of our data set and the other two public datasets are all used for testing. We use the "mean Average Precision (mAP)", which is commonly used for target detection literatures [8, 23], as our evaluation metric of the detection accuracy.

TABLE I
STATISTICS OF THE THREE DATASETS FOR AIRPLANE DETECTION: 1) OUR DATASET, 2) NWPU10 [23], AND 3) OCCLUDED DATASET [13]

Dataset name	Our Dataset	NWPU10 [23]	Dataset [13]
Resolution (m/pxl)	0.2-1.0	0.5-2.0	0.3-0.6
# images	21,161	90	47
# airplanes	2,005	757	184
Image size (avg.)	600×800	684×974	714×1,060

B. Implementation Details

The backbone of our network is built based on VGG16 [24] by removing "pool5" and its subsequent layers. We select "conv3_3", "conv4_3" and "conv5_3" to construct a three-layer feature pyramid. We set the size of anchor boxes as 36×36 on "feat3_p", 45×45, 62×62 and 81×81 on "feat4_p", 100×100, 140×140, 180×180 and 240×240 on "feat5_p".

As the number of targets and background windows are highly imbalanced in remote sensing images, the strategy of "Hard Negative Mining (HNM)" is used, i.e. only take the top 2% largest loss values of negative samples into consideration and back-propagate their gradients. The HNM is applied in both of the attention map prediction loss and the anchor category classification loss.

All images are resized to 600×800 in our experiment. During training, the anchors whose intersection over union with any ground truth box higher than 0.7 are considered as positive ones, and otherwise as negative ones. All training images are augmented by making random flip or rotations of 90, 180 and 270 degrees. We do not use any specific parameters that are associated with the orientation since improving rotation invariance is not the focus of this paper. We set $\gamma = 5.0$. As for the weights of different losses, we set $\beta_1 = \beta_2 = \beta_3 = 1$. In fact, we first tried to balance the losses to the same magnitude, but we observed the average precision decreases by 4.4% compared with using equal balance parameters.

We use momentum stochastic gradient descent for optimization. The initial learning rate is set to 0.001 and then decreases to its 1/10 every 10 epochs. The momentum is set to 0.005. The training stops when reaching its 50 epochs.

C. Detection Statistics and Ablation Analysis

The ablation analyses are made to analyze the importance of each technical component of our method, including 1) anchor design, 2) hard negative mining and 3) attention mechanism. The baseline methods are first evaluated, then we gradually integrate these techniques. The detection results on the three datasets are displayed in Table II. The first line shows the results of our baseline method. We can observe the improvement of the accuracy when adding these components on top of each other, which suggests all these strategies can help improve detection accuracy.

We also visualize the detection results on input images so that to figure out why the local attention branches are useful for recalling occluded airplanes. In Fig. 4, column (a) shows the detection results without using local attention, column (b) shows the detection results with local attention, and column (c) shows the predicted score of the attention branch. As we can see, for those occluded targets, their visible key points can also be recognized well by our method.

TABLE II
ABLATION STUDIES ON THREE TECHNIQUES 1) ANCHORS (ACHR), 2) HARD NEGATIVE MINING (HNM), AND 3) ATTENTION (ATT)

Ablations			Average Precision on Different Datasets		
ACHR	HNM	ATT	Our Dataset	NWPU10 [23]	Dataset [13]
			93.8%	87.0%	76.0%
✓			93.0%	90.3%	79.4%
✓	✓		95.4%	97.1%	83.3%
✓	✓	✓	**95.9%**	**98.2%**	**85.7%**

TABLE III
COMPARISON WITH OTHER STATE OF THE ART DETECTION ARCHITECTURES.

	Average Precision on Different Datasets		
Methods	Our Dataset	NWPU10[23]	Dataset[13]
Faster R-CNN [12]	97.0%	92.7%	78.2%
SSD [20]	**97.4%**	96.0%	78.9%
RFB-SSD [21]	97.1%	96.5%	79.1%
RefineDet [22]	96.2%	97.2%	81.7%
Ours	95.9%	**98.2%**	**85.7%**

To make more comprehensive evaluation of our method, we also compare with some SOTA detection systems, such as Faster R-CNN [12], SSD [20], RFB-SSD [21] and RefineDet [22]. Their results are listed in Table III, from which we can see that our method is competitive on our test set and performs better on the other two test sets, especially for the occluded target detection dataset [13]. As for the detection speed, we test our method on a 1080Ti GPU and PyTorch 0.3.1. The proposed method "RetinaNet+LAN" runs at 15.2fps, which has comparable detection speed with its baseline RetinaNet, which runs at 17.2fps.

IV. CONCLUSION

We propose a new detection method called local attention networks to improve the detection of occluded airplane targets in remote sensing images. The proposed method is designed with the idea of detecting by parts and is further integrated with an additional attention branch in the traditional detection pipeline so that to enhance the features of occluded targets and suppress those of the backgrounds. We conduct our experiments on three detection datasets. Detection results and ablation studies have demonstrated the effectiveness of our method, especially for those occluded airplane targets.

REFERENCES

[1] Y. Lecun, Y. Bengio, G. Hinton. "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436, 2015.
[2] Z. Li, L. Itti. "Saliency and gist features for target detection in satellite images," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2017-2029, 2011.
[3] W. Li, S. Xiang, H. Wang, *et al*. "Robust airplane detection in satellite images," in *IEEE ICIP*, 2011.
[4] Z. An, Z. Shi, X. Teng, *et al*. "An automated airplane detection system for large panchromatic image with high spatial resolution," *Optik*, vol. 125, no. 12, pp. 2768-2775, 2014.

(a) (b) (c)

Fig. 4. (Better viewed in color) Comparison of the detection results on dataset [13]: (a) without the using local attention (the yellow ones are missed but detected when using local attention), (b) with local attention, and (c) visualization of predicted score of the attention branch.

[5] Y. Zhang, Y. Zhuang, F. Bi, *et al*. "M-FCN: effective fully convolutional network-based airplane detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 99, pp. 1-5, 2017.
[6] F. Zhang, B. Du, L. Zhang, *et al*. "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7405-7415, 2016.
[7] G. Cheng, P. Zhou, and J. Han. "RIFD-CNN: rotation-invariant and fisher discriminative convolutional neural networks for object detection," in *IEEE CVPR*, 2016.
[8] Z. Zou and Z. Shi. "Random access memories: a new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100-1111, 2018.
[9] Cheng, Gong, Junwei Han, Peicheng Zhou, and Dong Xu. "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol 28, no. 1, pp. 265-278, 2019.
[10] Cheng, Gong, Peicheng Zhou, and Junwei Han. "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, 54, no. 12, pp. 7405-7415, 2016.
[11] R. Girshick, J. Donahue, T. Darrell. "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE CVPR*, 2014.
[12] S. Ren, K. He, R. Girshick, *et al*. "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137, 2017.
[13] S. Qiu, G. Wen, Z. Deng, *et al*. "Occluded object detection in high-resolution remote sensing images using partial configuration object model," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 5, pp. 1909-1925, 2017.
[14] S. Qiu, G. Wen, Y. Fan. "Automatic and fast PCM generation for occluded object detection in high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 99, pp. 1-5, 2017.
[15] P. Felzenszwalb, R. Girshick, D. Mcallester, *et al*. "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627-1645, 2010.
[16] T. Lin, P. Dollar, R. Girshick, *et al*. "Feature pyramid networks for object detection," in *IEEE CVPR*, 2017.
[17] T. Lin, P. Goyal, R. Girshick, *et al*. "Focal loss for dense object detection," in *IEEE ICCV*, 2017.
[18] R. Girshick. "Fast R-CNN," in *IEEE ICCV*, 2015.
[19] A. Shrivastava, A. Gupta, R. Girshick. "Training Region-based Object Detectors with Online Hard Example Mining," in *IEEE CVPR*, 2016.
[20] W. Liu, D. Anguelov, D. Erhan, *et al*. "SSD: Single Shot MultiBox Detector," in *ECCV*, 2016.
[21] S. Liu, D. Huang, Y. Wang. "Receptive Field Block Net for Accurate and Fast Object Detection," in *IEEE CVPR*, 2017.
[22] S. Zhang,L. Wen, X. Bian, *et al*. "Single-Shot Refinement Neural Network for Object Detection," in *IEEE CVPR*, 2018.
[23] G. Cheng, and J. Han. "A survey on object detection in optical remote sensing images," in *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11-28, 2016.
[24] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556v6, 2015.
[25] C. Stubenrauch, W. Rossow, and S. Kinne, et al., Assessment of global cloud datasets from satellites: Project and database initiated by the gewex radiation panel, *Bull. Am. Meteorol. Soc.*, vol. 94,no. 7, pp. 10311049, 2013.
[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.