Airplane Detection in Optical Remote Sensing Video Using Spatial and Temporal Features

1st Jing Bai
School of Artificial Intelligence
Xidian University
Xi'an, China
baijing@mail.xidian.edu.cn

2nd Wentao Yu
School of Artificial Intelligence
Xidian University
Xi'an, China
wentao.yu@stu.xidian.edu.cn

3rd Anran Yuan
School of Artificial Intelligence
Xidian University
Xi'an, China
aryuan@stu.xidian.edu.cn

4th Zhu Xiao
College of Computer Science and Electronic Engineering
Hunan University
Changsha, China
zhxiao@hnu.edu.cn

***Abstract*—Benefited from the rapid development of deep learning, object detection in natural image has made great improvements. However, since the size of the optical remote sensing video is very large while the size of airplane is very small, airplane detection in optical remote sensing video still faces a lot of challenges. In this article, we aim at a novel approach for airplane detection in optical remote sensing video. The proposed approach utilizes spatial features from structured forests edge detection and temporal features from neighboring frames. It is capable of circumventing existing challenges and running at a high speed for practical applications. To realize this goal, edge detection results of optical remote sensing video frames are obtained from structured forests edge detection method. Afterwards, improved frames differencing method is utilized to extract temporal features. Finally, airplane detection result is generated by deep neural networks with extracted spatial and temporal features. Our experiments demonstrate that our method has a great breakthrough on the precision and recall of airplane detection in optical remote sensing video.**

***Index Terms*—airplane detection, optical remote sensing video, structured forests edge detection, frames differencing**

I. Introduction

Airplane detection in optical remote sensing video is crucial in the aspect of military and civilian fields, such as air defense and airport surveillance. It attracts more and more attention [1]–[4]. With the increasing demand of airplane detection in optical remote sensing images, lots of methods have been proposed for airplane detection in recent years [5]–[7]. In these methods, a large part of the literature has focused on deep neural network (DNN), such as faster R-CNN [8], SSD [9] and YOLO [10].

Yun Ren *et al.* [11] integrate deformable convolution into the faster R-CNN [8]. Besides, they add 2D offsets to standard convolution layer and adopt top-down and skip connections for object detection in optical remote sensing images. Zhong Chen *et al.* [12] design an end-to-end deep learning method for airplane detection using the SSD detection model [9]. Heng Zhang *et al.* [13] use modified YOLO object detection model [10], [14] and feature extraction model intergrated with a self-designed layer named VaryBlock. These methods tap into the advantages of DNN's properties and become a promising trend for airplane detection.

Although existing research for airplane detection in optical remote sensing images has made promising progress, it is still facing several technical challenges in optical remote sensing video. The first technical challenge is that the resolution of optical remote sensing video is 12000×5000 while the size of airplane is less than 70×70. It results in a low contrast between the background and the target. The second technical challenge is that the spatial resolution of optical remote sensing video is less than $1m$, which results in poor motion features of airplanes.

To cope with these technical challenges, we propose a novel airplane detection method using spatial features from structured forests edge detection [15] method and temporal features from improved frames differencing, with the purpose of achieving a reliable and efficient performance in optical remote sensing video. Firstly, we utilize structured forests edge detection method to obtain spatial features of optical remote sensing video. Then, by taking the advantage of improved frames differencing method, temporal features are generated from adjacent frames. Finally, airplane detection result is obtained by deep neural networks with extracted spatial and temporal features.

The threefold contributions of our work can be outlined as follows.

- Our proposed method focuses on the most significant and unique structured forests edges in optical remote sensing video to extract spatial features. Structured forests edges can highlight the contrast between the airplane and the background, so as to greatly solve the technical challenge

This work was supported in part by the State Key Program of National Natural Science of China (No. 61836009), in part by the National Natural Science Foundation of China (No. 61772401), in part by the Open Fund of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education (Grant No. IPIU2019007).

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

that the contrast between the background and the target is too low.

- Our proposed method extracts temporal features from adjacent frames in optical remote sensing video. The result of frames differencing can strengthen the motion features of airplanes. It makes great contributions to alleviate the disadvantage of poor motion features of airplanes.
- We combine spatial features with temporal features by creating deep neural network named as fusion network. Extensive experiments are performed on the optiacl remote sensing video dataset whih is made by ourselves. Experimental results demonstrate that our method can achieve 0.9692 precision and 0.9333 recall on our test set, which is superior to the state-of-the-art methods.

II. Methodology

In this section, we present the methodology of the proposed airplane detection in optical remote sensing video. As shown in Figure 1, the framework of our method is made up of three parts. The first part is structured forests edge detection, which is aimed at obtaining spatial features of optical remote sensing video. The second part is improved frames differencing method for the purpose of extracting temporal features of optical remote sensing video. Afterwards, we concatenate extracted spatial and temporal features and feed them into deep neural networks. At last, airplane detection result in optical remote sensing video is obtained.

A. Structured Forests Edge Detection

First of all, we cut the optical remote sensing video into frames. Then, edge detection results of optical remote sensing video are obtained by structured forests edge detection. Due to the characteristic of optical remote sensing video, the edges of edge detection result are very tight. As a result, non-maximum suppression (NMS) algorithm is needed to improve it. The NMS algorithm suppresses edge points that are not maximum values and searches for local maximum values. It is a widely used algorithm in visual tasks such as edge detection and object detection. After NMS processing, the original edge image can be changed to a relatively sparse edge image, which is more conducive to subsequent operations.

Secondly, we utilize the grouping strategy. After grouping operation, the original edge detection results become colorful and discrete. Edge points are grouped into many short line segments. It can be seen from the third picture of Figure 2 vividly. Our grouping strategy is to form an edge set by taking edge points near the same line in the edge detection result as members of the same group. The criterion is to keep finding 8 connected edge points until the sum of the difference in the direction angle between each edge point is greater than $\frac{\pi}{2}$.

B. Improved Frames Differencing

The traditional two-frame differencing method [16] is to subtract the corresponding pixel of the $n-1_{th}$ frame from the n_{th} frame. After the two-frame differencing image is obtained, the absolute value of grayscale difference is determined. When the absolute value is higher than a certain threshold, it is determined to be the moving target. The three-frame differencing method is a variant of the two-frame differencing method. Obtain two difference images between three adjacent frames at first. Then perform an AND operation on the pixels at the same position of two difference images to obtain the final three-frame differencing result.

For optical remote sensing video with very high resolution, the size of the target is small compared to the overall scene. Besides, the relative position difference of the moving target between adjacent frames is small, which leads to the relative displacement of airplanes is small. If the two-frame differencing method or its variant method is used, the frames differencing result cannot extract sufficient temporal features. Therefore, we need to improve the traditional frames differencing method to extract better temporal features.

The steps of the improved frames differencing method are as follows.

Algorithm 1: Improved frames differencing

Result: frames differencing result

cut the optical remote sensing video into N frames denoted by F;
perform spatial domain based image enhancement operation;
extract frames every three frames and reconstitute them into a sequence S with length M, where $N = 3 \times M$;
$i \leftarrow 0$;
while $i < M$ **do**
    subtract the corresponding pixels of F_{3i-1} from F_{3i} and denote the result as D ;
    $i \leftarrow i + 1$;
end
threshold segmentation;
morphological expansion operation

As is shown in the first picture of Figure 3, this original frame is extracted from the optical remote sensing video. It is the $270th$ frame in the video. The second picture of Figure 3 is the result after spatial domain based image enhancement operation. Then subtract the corresponding pixels of $269th$ from $270th$. After subtraction, threshold segmentation is utilized to obtain the third picture of Figure 3. The last picture of Figure 3 is the result after morphological expansion operation. The size of morphological expansion operator in our work is 2×2. The void phenomenon caused by frames differencing method can be effectively suppressed by expansion operation.

C. Deep Neural Networks

1) Fusion Network: In the application of airplane detection in optical remote sensing video, it is required to extract features more efficiently by stronger generalization capability. Therefore, we feed structured forests edge detection result and improved frames differencing result into a convolutional neural network (CNN) before airplane detection. We name this CNN as fusion network. It is modified from VGG-13 [17]. The

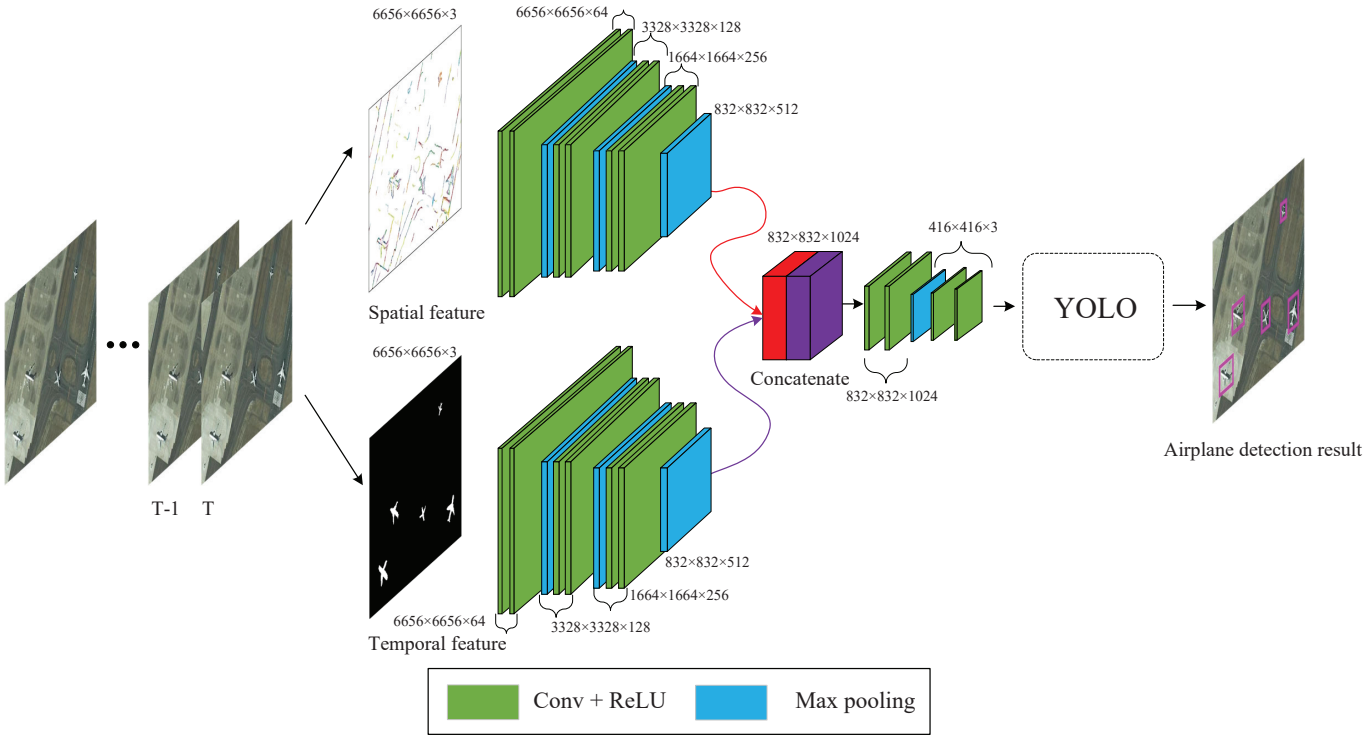

Fig. 1. Airplane detection in optical remote sensing video using spatial and temporal features

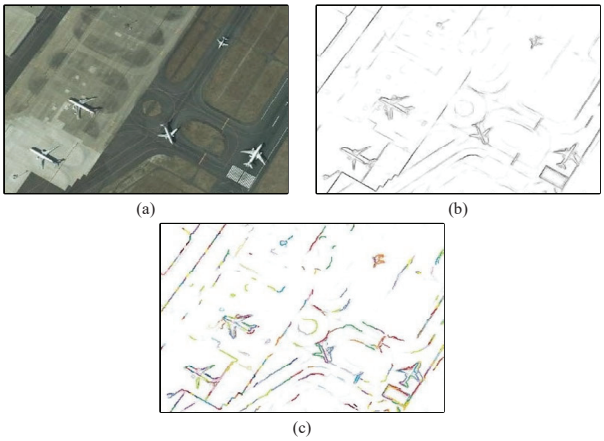

Fig. 2. Structured forests edge detection: (a) original optical remote sensing image (b) the result of structured forests edge detection (c) edge set after using grouping strategy

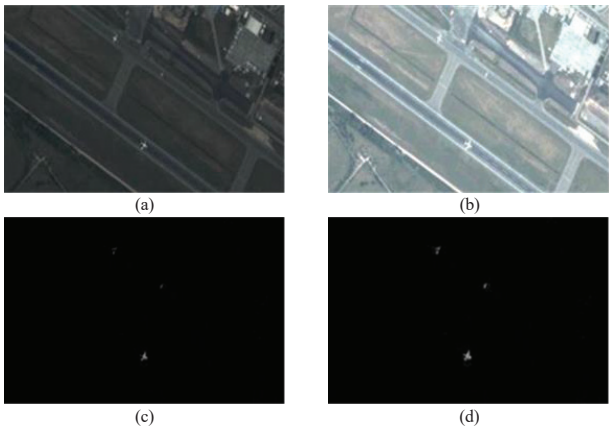

Fig. 3. Improved frames differencing: (a) original frame in optical remote sensing video (b) after spatial domain based image enhancement operation (c) after subtraction operation and threshold segmentation (d) after morphological expansion operation

fusion network is divided into two parts. The first part is made up of 6 convolutional layers and 3 max pooling layers. The second part consists of 4 convolutional layers and one max pooling layer. The complete network is depicted in the Figure 1 in detail.

First of all, the results of structured forests edge detection and improved frames differencing are resized to $6656 \times 6656 \times 3$. Then, they are fed into two subnetworks with 6 convolutional layers and 3 max pooling layers respectively. When we obtain two $832 \times 832 \times 512$ feature maps, we concatenate these feature maps sequentially and achieve a $832 \times 832 \times 1024$ feature map. Afterward, it is fed into the second part of our fusion network.

In the second part of our fusion network, the max pooling layer is different from the original one in VGG-13. We reduce the dimension of feature maps in subsequent layers in order to fit the input dimension of YOLO detection network.

2) Detection Network: Due to the outstanding performance of YOLO, we choose it as our detection network. The architecture of YOLO is shown in Figure 4. It is made up of

22 convolutional layers, 5 maximum pooling layers, and one passthrough layer. The input of the network is $416 \times 416 \times 3$.

i) The input image passes through the first convolutional layer with a convolution kernel of 3×3, and then passes through a max pooling layer to obtain 32 feature maps with a size of 208×208.

ii) It passes the second convolutional layer with 3×3 convolutional kernel size. After that, 64 feature maps with the size of 104×104 are obtained through a max pooling layer. 128 feature maps with the size of 52×52 are obtained by continuously passing through three convolutional kernel size of 3×3, 1×1, 3×3 and one max pooling layer.

iii) Three convolutional layers with 3×3, 1×1, 3×3 and one max pooling layer are successively passed through to obtain 256 feature maps with a size of 26×26. 512 feature maps with dimensions of 13×13 are obtained by continuously passing through the five convolutional layers with sizes of 3×3, 1×1, 3×3, 1×1 and 3×3 and one max pooling layer.

iv) Through seven convolutional layers with sizes of 3×3, 1×1, 3×3, 1×1, 3×3, 3×3, and 3×3, 512 feature maps with sizes of 26×26 are obtained.

v) Through the convolutional layer with convolution kernel size of 1×1, the feature map is rearranged through the passthrough layer. 256 feature maps with 13×13 size are obtained. At last, through the convolution kernel sizes of 3×3 and 1×1, a $13 \times 13 \times 30$ feature map is obtained, which can generate the final airplane detection result of optical remote sensing video.

III. Experimental Results And Analysis

Extensive experiments are performed on the optiacl remote sensing video dataset, which is made by ourselves to evaluate the performance of the proposed airplane detection method. Then we compare our proposed airplane detection method with other advanced methods.

A. Dataset

Our dataset is based on the high-resolution optical remote sensing videos provided by Jilin No. 1 satellite. The videos are in AVI format and the resolution of videos are less than $1m$. The ground area covered by each video is $11km \times 4.5km$. All the videos have undergone geometric correction, radiometric correction and image stabilization. The main targets in our optical remote sensing videos are fixed-wing airplane. The average length of each video is 30 seconds. Figure 5 shows one of optical remote sensing video frames at Bogota's airport.

Since the resolution of optical remote sensing videos is 12000×5000 including a lot of interference such as background clutter and moving cloud occlusion, we preprocess the videos at first.

Firstly, we cut out the cloud covered fragments and only keep the first 17 seconds of the video. Figure 6 shows the first, $101th$ and $201th$ frame respectively.

Secondly, in order to make the video frames more suitable for the structure of deep neural network, we extract 10 image block samples from the part containing the airplane, which are depicted in Figure 7. The number of airplanes contained in each image ranges from 3 to 9, and the total number of airplanes contained in our dataset is 680. In the test set of our dataset, there are 20 video frames with 135 airplanes.

B. Improved Frames Differencing

Figure 8 depicts the results of frames differencing methods. As can be seen from Figure 8, when using the traditional two-frame differencing method to detect moving airplanes in high-resolution optical remote sensing videos, only the contours of the airplanes moving from left to right can be detected, which results in void phenomenon. In addition, for smaller airplanes moving from right to left, the detection results are less clear. In conclusion, the temporal features extracted from traditional two-frame differencing method can not satisfy real demands.

The proposed improved frames differencing method makes inter-frame time interval selection more appropriate, which effectively suppresses the void phenomenon. The airplane moving from left to right in the fifth and sixth picture of Figure 8 is almost completely detected and the smaller target from right to left can be seen in shape and position. It demonstrates that the proposed improved frames differencing method is effective. What's more, the proposed method has the advantages of simple implementation, low design complexity and high stability, which achieves very good performance in high-resolution optical remote sensing videos.

C. Experimental Parameter Settings

In the improved frames differencing method, the block size of the expansion operation is 2×2. The frame interval of the differential image is set to 3. In our deep neural networks, the learning rate is set to 0.001, the size of each training batch is set to 64, the maximum number of training iterations is set to 30000, and the network weights is automatically saved every 1000 epochs. The rates of saturation and exposure are set to 1.5. The target detection probability threshold is set to 0.1, that is, the airplane is determined when the probability exceeds 0.1.

D. Experiments

According to the set parameters, we train our deep neural networks. Figure 9 shows the broken line graph of loss and average IoU with iteration times. According to the results in the Figure 9, after comprehensively considering the higher average IoU and the lower loss, the weight parameter after the $19000th$ iteration is selected for verifying the performance of our proposed airplane detection method.

We perform experiments on 20 video frames in test set. Airplane detection results are depicted in Figure 10. The red bounding boxes in Figure 10 are detected airplanes.

Our analysis of the experimental results is as follows. Among 135 airplanes in the test set, 126 airplanes are successfully detected (True Positive) and 9 airplanes are missed(False Negative). In addition, 4 non-targets are misidentified as airplanes (False Positive). From this data, the detection precision, recall and F-Measure of the proposed method can be calculated according to the following formulas.

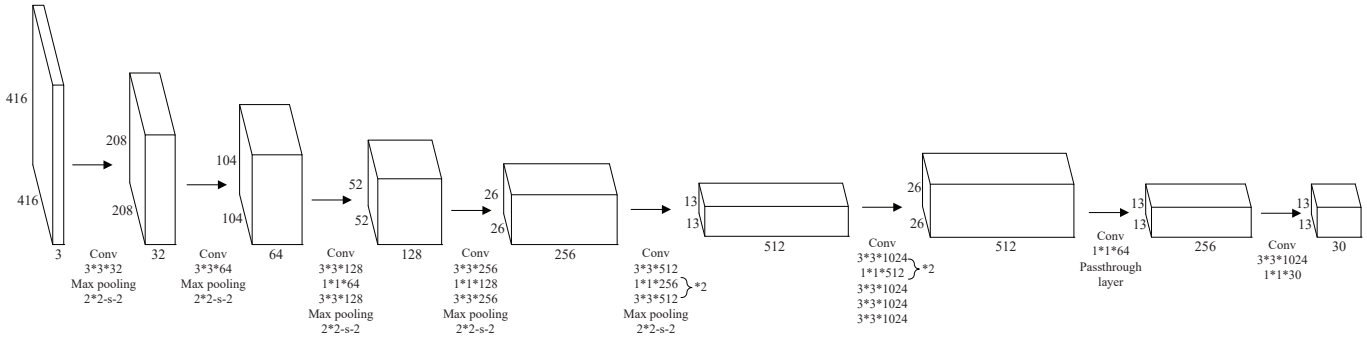

Fig. 4. The architecture of YOLO

Fig. 5. One of optical remote sensing video frames at Bogota's airport

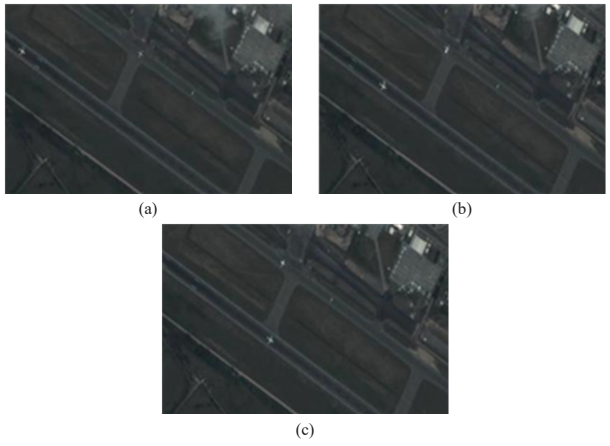

Fig. 6. Cropped video:(a) the first frame (b) the 101th frame (c) the 201th frame

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

After calculation, the precision of the proposed method is 0.9692, the recall of the proposed method is 0.9333 and the F-Measure of the proposed method is 0.9509.

According to the precision, recall and F-Measure, the proposed airplane detection method has obtained excellent performance, which is very competitive in high-resolution optical remote sensing video for airplane detection.

The comparative experimental method chosen in this article is R-CNN and BOVW [18]. Table 1 illustrates the comparison results of different airplane detection methods on precision, recall and F-Measure. As can be seen from Table 1, our proposed method performs better than the other two methods, which has a significant improvement. It is benefited from extracted spatial and temporal features and powerful deep neural networks.

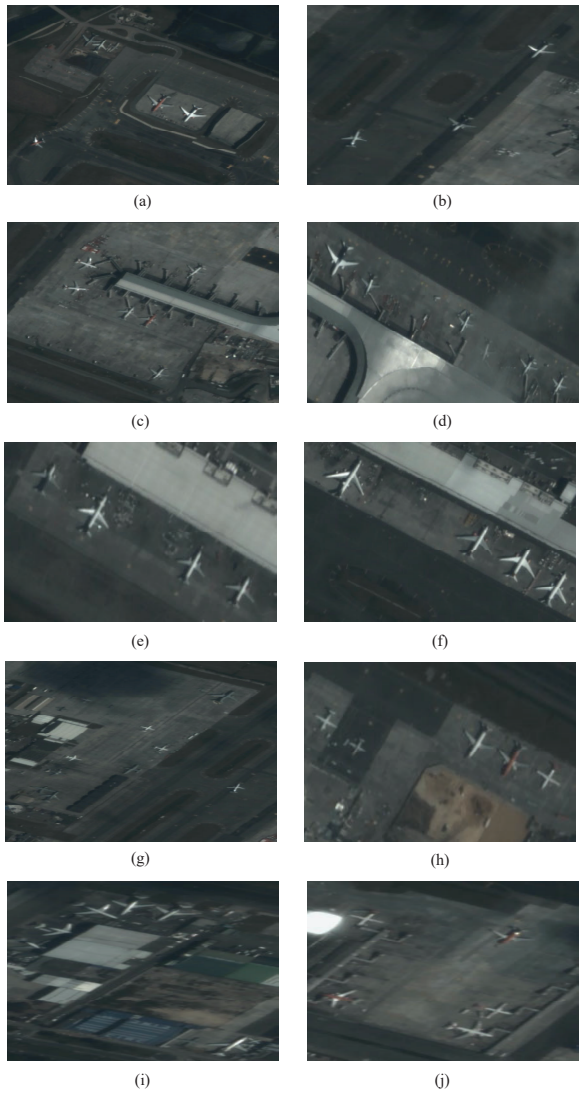

Fig. 7. 10 image block samples with airplanes

TABLE I
COMPARISON OF DIFFERENT AIRPLANE DETECTION METHODS

Method	BOVW	R-CNN	Proposed
Precision	0.265	0.684	0.969
Recall	0.632	0.796	0.933
F-Measure	0.373	0.736	0.951

To sum up, our proposed method can effectively implement airplane detection in high-resolution optical remote sensing video.

IV. CONCLUSION

In this article, we develop a robust airplane detection method via using spatial and temporal features. The proposed method utilizes structured forests edge detection method to obtain spatial features of optical remote sensing video. In addition, by taking the advantage of improved frames differencing method, temporal features are generated from adjacent frames.

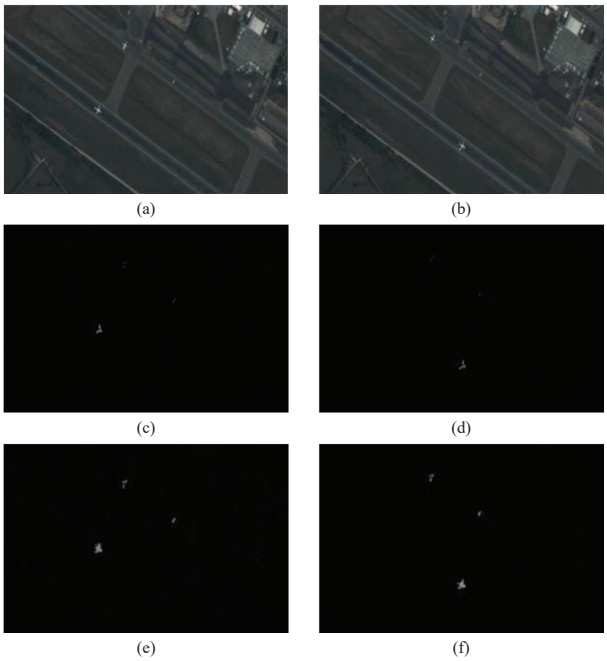

Fig. 8. The results of frames differencing methods: (a) The 150th frame in the optical remote sensing video, (b) The 270th frame in the optical remote sensing video, (c) The result of the 150th frame using traditional two-frame differencing method, (d) The result of the 270th frame using traditional two-frame differencing method, (e) The result of the 150th frame using the proposed improved frames differencing method, (f) The result of the 270th frame using the proposed improved frames differencing method.

Fig. 9. Broken line graph of loss and average IoU with iteration times

Airplane detection result is obtained by deep neural networks with extracted spatial and temporal features at last. Extensive experiments are performed on the optiacl remote sensing video dataset, which is made by ourselves. The results domonstrate that the proposed method achieves 0.9509 F-Measure on our test set, which obtains the state-of-the-art performance on airplane detection in optical remote sensing video.

It is noteworthy that the quality and quantity of optical remote sensing video datasets for airplane detection used in academia are limited. In the future, we will devote our effort to collecting more optical remote sensing video datasets for airplane detection to support in-depth research.

REFERENCES

[1] Q. Luo and Z. Shi, "Airplane detection in remote sensing images based on object proposal," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2016, pp. 1388–1391.

Fig. 10. Airplane detection results in optical remote sensing video

[2] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016.

[3] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[4] L. Cao, "Fine-grained road mining from satellite images with bilateral xception and deeplab," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[5] K. L. de Jong and A. S. Bosman, "Unsupervised change detection in satellite images using convolutional neural networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[6] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2014.

[7] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] Y. Ren, C. Zhu, and S. Xiao, "Deformable faster r-cnn with aggregating multi-layer features for partially occluded object detection in optical remote sensing images," *Remote Sensing*, vol. 10, no. 9, p. 1470, 2018.

[12] Z. Chen, T. Zhang, and C. Ouyang, "End-to-end airplane detection using transfer learning in remote sensing images," *Remote Sensing*, vol. 10, no. 1, p. 139, 2018.

[13] H. Zhang, J. Wu, Y. Liu, and J. Yu, "Varyblock: A novel approach for object detection in remote sensed images," *Sensors*, vol. 19, no. 23, p. 5284, 2019.

[14] W. Zhihuan, C. Xiangning, G. Yongming, and L. Yuntao, "Rapid target detection in high resolution remote sensing images using yolo model," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, p. 3, 2018.

[15] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1841–1848.

[16] X. Wang, L. Liu, G. Li, X. Dong, P. Zhao, and X. Feng, "Background subtraction on depth videos with convolutional neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.