IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 9, 2019, accepted September 27, 2019, date of current version October 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2945834

Improvement of Non-Maximum Suppression in RGB-D Object Detection

DECHENG WANG, XIANGNING CHEN, HUI YI, AND FENG ZHAO
Academic of Space Information, University of Space Engineering, Beijing 101416, China

Corresponding author: Decheng Wang (wangdecheng@tom.com)

This work was supported in part by the National Defense Science and Technology Innovation Zone Special Program under Grant 18-H863-01-ZT-002-055.

ABSTRACT Currently, the non-maximum suppression (NMS) algorithm is a commonly used method in the post-processing stage of object detection. However, the NMS algorithm cannot effectively eliminate missing and false object detection results because of the simple constraint condition. To solve the problem of the poor detection effect in highly overlapping dense object scenes in the traditional NMS algorithm, we design an RGB-D object detection network model based on the YOLO v3 framework, and using level-by-level metaphase fusion on the RGB and depth information, we propose an improved NMS algorithm which fuses depth characteristics. According to the depth of the object in the detection boxes, it is determined whether another object is the same object in highly overlapping detection boxes, and the average depth of the internal pixels in the detection boxes is calculated as a penalty term, then the penalty term is added to the detection box score to obtain a new constraint condition for non-maximum suppression. The experimental results on the NYU Depth V2 dataset show that the mean average precision (mAP) of the Depth Fusion NMS algorithm proposed in this paper is 0.8%, 0.5% and 0.3% higher than those of the Greedy-NMS, Soft NMS-L and Soft NMS-G methods, respectively. After comparison and analysis, our method can not only detect more overlapping objects but also achieve a better object localization accuracy.

INDEX TERMS Non-maximum suppression, RGB-D object detection, intersection-over-union, detection boxes, multimodal fusion.

I. INTRODUCTION

Object detection is an important research direction in the field of computer vision. The process can be understood as visual algorithm giving the computer a human-like visual recognition ability to identify object categories and obtain the object location information in scenes through an image obtained by a sensor. In recent years, with the rapid development of deep learning and neural network technology, the research on object detection has resulted in breakthroughs in the areas of monitoring security, automatic driving, human-computer interaction and so on [1]. Object detection algorithms based on convolutional neural networks can be divided into three steps [2]: feature learning and object extraction, object classification and location regression, and non-maximum suppression algorithms to select the optimal detection boxes. Non-maximum suppression (NMS) in the last step was first proposed in the edge detection algorithm, and then further applied to the fields of object detection, face recognition, etc. [3], [4]. NMS is an important method for the post-processing step of a detection model. Current studies mainly focus on feature learning, feature extraction and classification, but there has been little improvement in non-maximum suppression algorithms [5].

With the popularity of consumer-level depth sensors (such as Kinect), we can easily obtain the depth information of objects in a scene, which greatly promotes the application of RGB-D images in related fields such as object detection. The gray scale value of each pixel in a depth image represents the distance from the corresponding object in an RGB image to the camera. References [6]–[9] and other papers have shown that adding one-dimensional depth information to an RGB network can effectively avoid the impact of illumination changes and other factors for object detection results, which can improve the accuracy and recall rate of detection model. However, RGB-D object detection methods based on

The associate editor coordinating the review of this manuscript and approving it for publication was Jingchang Huang.

convolutional neural networks (CNNs) mostly research the fusion of RGB and depth features and the network structure. The traditional NMS algorithm is still used in the network post-processing stage to select optimal detection boxes by comparing the prediction score and the size relationship between the IoU value of overlapping detection boxes and a given threshold T. However, the selection of the threshold T is usually determined through experience, which is likely to cause instability in the system detection accuracy. In view of the above problems, this paper improves the NMS algorithm for RGB-D object detection, adjusts the detection box score by using the depth characteristics of different objects, and obtains the optimal detection boxes for each object, thereby effectively reducing the false and missing detection rate of the detection model. In this paper, we applied the improved NMS algorithm in the current, popular detection framework YOLO v3 [10], and the network model was trained and tested in the challenging RGB-D dataset NYU Depth V2 [11], then we obtained a high mean average precision (mAP).

II. RELATED WORK

The commonly used non-maximum suppression algorithm is a greedy strategy. Only single overlapping area information is used for suppression. To improve the algorithm accuracy in the post-processing stage of object detection, some researchers have made corresponding improvements to the NMS algorithm. In 2015, reference [5] combined the scale ratio, the detection score ratio and the peripheral window information in NMS algorithm based on the ACF (aggregate channel features), which significantly improved the accuracy of the algorithm but simultaneously increased the time consumption, and the algorithm is only improved for pedestrian detection, lacking versatility.

In 2016, aiming to solve the problem that the traditional NMS constraint condition is too simple to eliminate the overlapping detection efficiently, Zhang *et al.* [12] proposed an improved, simplified non-maximum suppression algorithm, which added ''completed covered detection suppression'' and ''PASCAL VOC overlap criterion'' constraints, which calculate the coverage ratio of the intersection area to the selected detection bounding box and the overlap ratio of the combined area, respectively. The experimental results show that the improved method can reduce the error and improve the detection performance, but it still has involves threshold selection and misses small objects.

An improved NMS method was proposed in reference [13] in 2017. A part of the NMS loss is added to the loss function of the network according to the NMS location error. The NMS loss is similar to the classification loss, and the NMS error can be continuously reduced by back propagation during network training. Although the detection accuracy can be improved in this way, the introduction of the NMS loss function leads to an increase in the training time of the network, and the network parameters are redundant, which is not conducive to lightening the weight of detection model.

In 2018, Qiu *et al.* [14] determined that the performance of the NMS algorithm is substantially affected by highly overlapping objects, and its localization accuracy only depends on the highest score detection. Therefore, they proposed an accurate NMS method, which gradually merges highly overlapping detection boxes in an iterative manner, taking advantage of Regression-NMS [15] and Soft NMS, while eliminating their disadvantages. The experimental results show that this method can not only detect more overlapping objects but can also achieve a better object localization accuracy. In the same year, Zhao et al. proposed an improved NMS algorithm in reference [2]. First, according to the IoU value of the detection box and the preselected detection box, the proportional penalty factor corresponding to the detection box is calculated; then detection box confidence score is multiplied by the proportional penalty factor, and the score of the detection box is reduced by the proportional penalty factor one by one; finally, after several iterations, the detection box whose score is lower than the threshold is removed. Experiments showed that the improved NMS algorithm can effectively preserve the object detection box and remove the false positive detection box, thus reducing the missing and false detection rate of the NMS algorithm. Both of these algorithms improve the detection accuracy in an iterative manner, but the iterative process not only increases the number of calculations and is time consuming but also it cannot solve the problem of missed detection of intensive objects with high overlapping.

Although the traditional NMS algorithm is used in the post-processing stages of popular object detection algorithms such as SSD [16], Faster R-CNN [17] and YOLO v3 [10] and achieves a good performance, it is still an obviously flawed greedy algorithm. This paper aims to improve the NMS algorithm in a double-channel RGB-D convolutional neural network by using object depth characteristics, effectively reducing the localization error of the detection box and decreasing the missing detection rate of highly overlapping intensive objects, thereby, improving the accuracy of the detection model.

III. ALGORITHM DESIGN

In this section, we introduce the principle of the traditional NMS algorithm and explain our improved NMS algorithm process in detail.

A. TRADITIONAL NMS ALGORITHM

Non-maximum suppression can be understood as a local maximum search, which has very important applications in the field of computer vision [18]. In object detection, the NMS algorithm is often used to extract the prediction box with the highest score. The process involves extracting the feature from the sliding window, and after the classifier recognizes the classification, each detection box receives a score, but the sliding window will yield many detection boxes containing or mostly intersecting other windows. Then, NMS is needed to extract the prediction boxes with the highest scores in the neighborhood (the probability that an object is

the largest) and suppress the prediction boxes that have other lower scores. The process of the non-maximum suppression algorithm is shown in Fig. 1.

FIGURE 1. The diagram of the non-maximum suppression algorithm.

The principle of NMS is not complicated, and mainly involves calculating the IoU of each overlapping detection box and comparing it with the threshold T to determine the final detection box. IoU refers to the ratio of the intersection and the union for two detection boxes areas (intersection-over-union), whose formula is described as follows:

$$IoU = \frac{area(BB_i \cap BB_j)}{area(BB_i \cup BB_j)} \tag{1}$$

where BB_i and BB_j are two different detection boxes and area indicates the detection box area. For the list of all detection boxes and their corresponding confidence values $\boldsymbol{S}$, first select the detection box M with the largest score, remove it from the collection $\boldsymbol{B}$ and add it to the final detection result $\boldsymbol{D}$, and then calculate the IoU value of M and remaining detection boxes in $\boldsymbol{B}$, which removes the box that is larger than a certain threshold T to form set $\boldsymbol{B}$. Repeat this process until it is empty. The specific steps are described as follows:

1). Sort the scores of all the detection boxes, then select the highest score and its corresponding box;

2). Scan the remaining detection boxes, if the overlapping area (IoU) with the current highest score is larger than threshold T, then delete the corresponding box;

3). Continue to select the detection box with the highest score from the unprocessed detection boxes and repeat the above process.

B. DEPTH FUSION NMS ALGORITHM

The non-maximum suppression algorithm is used in the post-processing stage of object detection and plays an important role in ensuring the accuracy of detection box localization. However, the traditional NMS has two obvious defects. First, the selection of the optimal detection box only depends on the prediction score, which lacks robustness. Second, two objects that are close together will not be detected at the same time, as shown in Fig. 2. Aiming to solve the above problems, we propose an NMS post-processing method based on depth fusion and the depth characteristics of RGB-D images to make some corresponding improvements. The goal is to improve the missing detection rate and localization accuracy by introducing deep fusion terms.

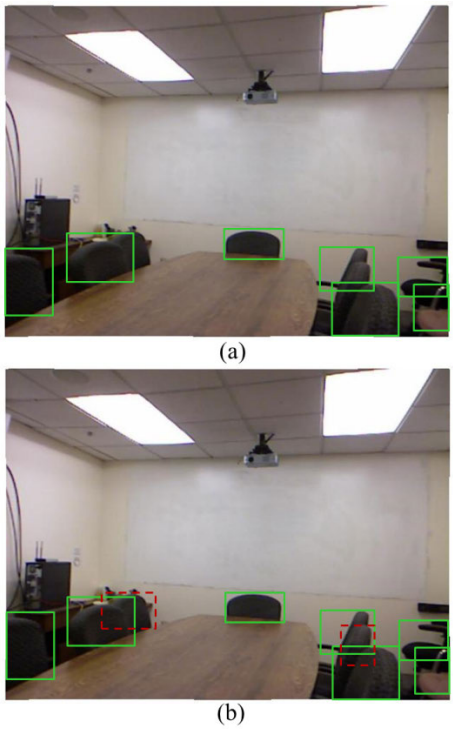


FIGURE 2. The detection results of "chair" in the NYU Depth V2 dataset: (a) is the result by traditional NMS (T_{IoU}= 0.6); (b) is the marked ground truth.

When using RGB-D images for object detection, we take YOLO v3 and Darknet-53 as the basic framework and network structure of the convolutional neural network, respectively. Inspired by the RGB-D network with level-by-level feature fusion proposed in [19], we design a double-channel network structure to extract RGB and depth features in the early stage, which integrates depth features into the branches of each scale feature in the middle of the RGB network to carry out the next forecast classification. Finally, in the post-processing stage, we propose an improved NMS method based on depth fusion. The overall network model structure is shown in Fig. 3.

For the feature fusion of RGB-D images, the most convenient method is to use the depth image content as the fourth channel of the RGB image, combine the two types of images or feature images into a four-channel image format, and then input them together into the convolutional neural network for feature extraction and object prediction. Another method is to extract the RGB and depth features simultaneously in two networks, and finally merge the features of the

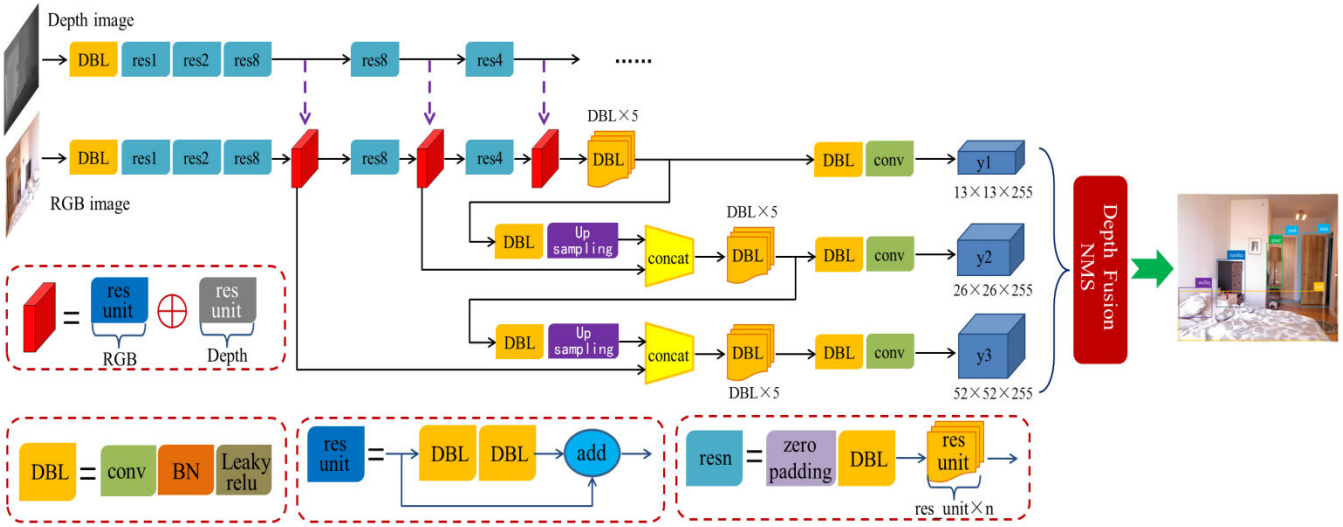


FIGURE 3. Structure diagram of the RGB-D object detection network model.

two modes in the fully connected layer. These two methods are common network structures for RGB-D object detection, but similar splicing can only learn a simple linear combination of RGB and depth information but cannot effectively explore the deeper correlation between the two modes. Therefore, the improvement in the detection effect after fusion is not obvious. In this paper, we propose an improved, two-channel network structure with level-by-level feature fusion. The correlation feature between RGB and depth mode is learned from the semantic feature expression of the middle layer. We use the RGB channel as the main network, and the depth network information is merged with the three scale feature layers of the main network, and the merged features are sent to the network branches of different scales for RGB-D object detection.

The fusion strategy of RGB and depth modes learns the correlation feature between the two modes by sharing weights, but the semantic information contained in the input feature maps $\mathbf{X}_{RGB}$ and $\mathbf{Y}_{\text{Depth}}$ is not completely equivalent; In order to more accurately fuse the two features, we use the "concatenate" feature fusion mode used in the DenseNet network [20] to effectively combine the two kinds of information. The "concatenate" operation is to extract the features of multiple convolution kernels or to fuse the information of the output layer. The fusion here refers to merge the number of feature channels, which increases the characteristics of the description image itself, and it is obviously beneficial for the classification of the final image. In the process of merging channels, we use the accuracy of the individual detection of RGB and Depth networks to determine the weight of two modal information fusion.

Assuming that the inputs of the RGB and depth channels are $\mathbf{x}_1, \mathbf{x}_1, \cdots \mathbf{x}_n$ and $\mathbf{y}_1, \mathbf{y}_1, \cdots \mathbf{y}_n$, respectively, the output of the combined channel is shown in (2), where α and β are the fusion weights of the RGB and depth features respectively, $\mathbf{W}_r$ and $\mathbf{W}_d$ are respectively weights by training two corresponding networks. ACC_{rgb} and ACC_{depth} are the accuracy of the RGB and depth images detection results, respectively, $\mathbf{Z}_{ri}$ and $\mathbf{Z}_{di}$ are the i neuron output of RGB and Depth networks.

$$\mathbf{Z}_{concat} = [\mathbf{Z}_{ri}; \mathbf{Z}_{di}] = \left[\alpha \sum_{i=1}^{n} \mathbf{x}_i * \mathbf{W}_{ri}; \beta \sum_{i=1}^{n} \mathbf{y}_i * \mathbf{W}_{di}\right] \tag{2}$$

$$\frac{\alpha}{\beta} = \frac{ACC_{\text{rgb}}}{ACC_{\text{depth}}} \tag{3}$$

$$\alpha + \beta = 1 \tag{4}$$

In the Depth Fusion NMS module of Fig. 3, we first judge the size relationship between the IoU value U of two overlapping detection boxes and the threshold T; if $U < T$, the detection box is retained; if $U \geq T$, the depth values of the center pixels of the two detection boxes in the depth image are compared. If there is a significant difference, then there are two objects in the two detection boxes. In this case, the two detection boxes should be preserved. If there is no significant difference, the objects in the two boxes belong to the same object. Then, we compare the scores S of the fused depth information, and the higher score is taken as the optimal detection box. The formula for S is as follows:

$$S_i = Score_i + \frac{1}{\ln(\overline{D_i})} \tag{5}$$

$$\overline{D_i} = \frac{1}{M_i \times N_i} \sum_{j \in \Omega_i} D_j \tag{6}$$

where $Score_i$ is the score of the ith detection box, $\overline{D_i}$ is the average gray value of the pixels in the ith detection box and represents the average depth, and M and N are the width and height, respectively, of the detection box. We can consider the depth value of the center pixel of a detection box as the approximate depth estimation of an object in the box. If the center pixel depth values of two detection boxes are similar

(less than the empirical value), then the objects detected by the two boxes are the same object. Since the average depth of pixel is smaller, the proportion of foreground objects in the detection box is larger, and the localization is more accurate. Therefore, the optimal solution is determined by combining the detection box score and the pixel average depth (as shown in (5)). The pseudocode for its process is described in Table 1.

TABLE 1. Pseudocode of the Depth Fusion NMS algorithm.

Input:	box set $\mathbf{R} = \{R_1, R_2, \ldots, R_n\}$ and its corresponding score $Score_i$, IoU threshold T, error empirical value ε, pixel depth value $\boldsymbol{D}$ of the depth images
1:	Set **B** = **R**
2:	**for** i=1: size(**R**) **do**
3:	**for** j=i+1: size(**R**) **do**
4:	Set U=IoU(B_i, B_j)
5:	**while** $U \geq T$ **do**
6:	**Define** D_{oi} and D_{oj} as the center pixel depth values of the two detection boxes.
7:	**while** $D_{oi} - D_{oj} < \varepsilon$ **do**
8:	Calculate the average depths $\overline{D_i}$ and $\overline{D_j}$ of the two detection boxes i and j according to (6)
9:	Calculate the scores S_i and S_j after fusing the depth of the two detection boxes i and j according to (5)
10:	**if** $S_i < S_j$ **then** remove B_i from **B** **end if**
11:	**end while**
12:	**end while**
13:	**end for**
14:	**end for**
15:	A new detection frame set **B** is obtained after Depth Fusion NMS;
Output:	**B**

IV. EXPERIMENT AND ANALYSIS

A. MODEL TRAINING PROCESS

We used the NYU Depth V2 RGB-D dataset to train the network and test performance of the improved algorithm. It is a challenging indoor scene classification database [11] established by New York University's Silberman et al., which contains 1,449 images of 464 different scenes, of which the RGB and depth image resolutions are both 640 × 480. The dataset has following characteristics: 1. The scene is photographed by a Microsoft Kinect v2 sensor, and the depth image is corrected by using a specific correction technique; 2. the shaded area of the depth image is well filtered by a cross bilateral filter, which can repair the depth image. 3. Adding a three-axis accelerometer to the Kinect camera eliminates the tilt and sloshing that occurs during sample collection. Some of the scenes in the dataset are shown in Fig. 4.

Experiments such as model training, feature fusion, objects detection and recognition were performed in Python 3.5 and run on GPU-accelerated drivers equipped with CUDA 9.0. The specific configuration of the experimental environment is shown in Table 2.

TABLE 2. Specific configuration of the experimental environment.

Configuration item	Version/Model
CPU	Intel(R) Xeon(R) CPU E5-2695 v2
GPU	NVIDIA GeForce GTX TITAN X
Memory	192 GB (@2.4 GHz)
Platform	Spyder Python 3.5

This chapter selects 1250 RGB-D images with different indoor scenes with completed depth from the NYU Depth V2 dataset, and 14 categories (tv, chair, desk, whiteboard, door, trash can, people, blackboard, cabinet, lamp, sofa, bed, phone and toilet) are used for training and testing, including 1000 training sets and 250 test sets. The experiment uses batch normalization, 64 pictures for training per iteration, and 30200 iterations. In the training stage, we use the stochastic gradient descent with a momentum term of 0.9. The initial learning rate of the weight is 0.001, and the decay coefficient is set to 0.0005. To better observe the training situation and

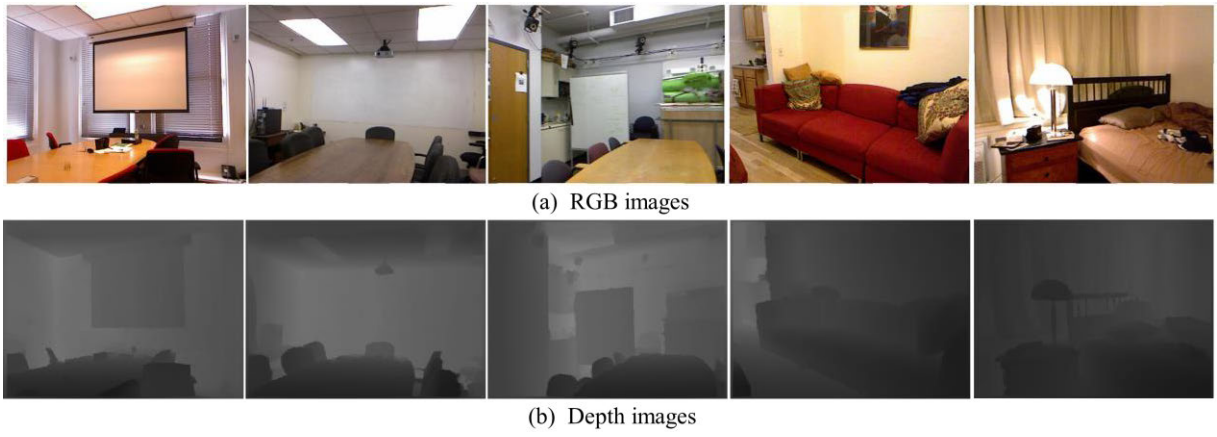


FIGURE 4. Some of the scenes in the NYU Depth V2 dataset.

evaluate the model performance, we introduce the loss function (loss), intersection-over-union (IoU) and recall rate into the training process, which are visualized in Fig. 5.

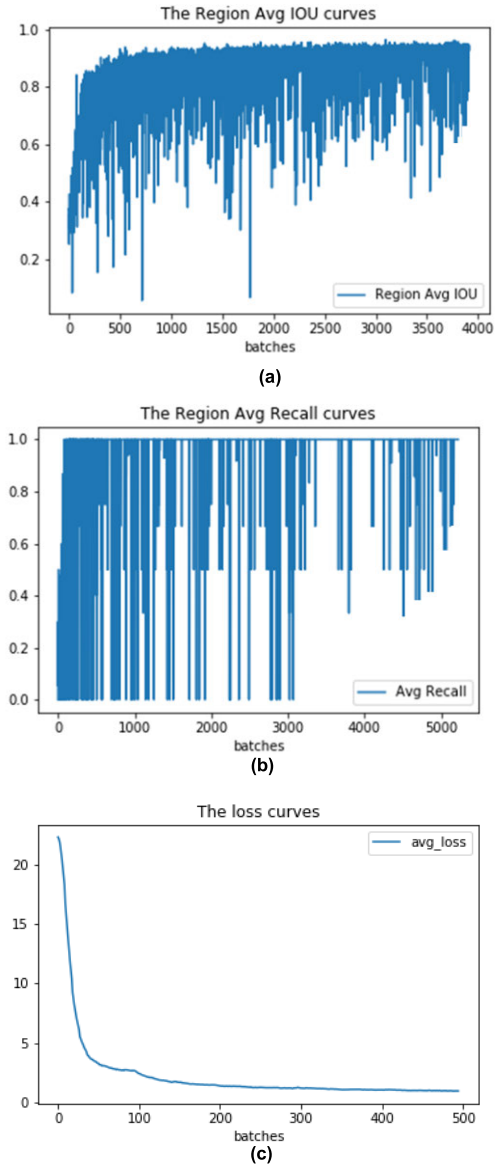


FIGURE 5. Changes of indicators during network training. (a) Shows the average IoU curve; (b) shows the average Recall curve; (c) is the loss curve for Iteration top 500 batches.

Since the number of iterations is large during the training process, we can observe the model training clearly after downsampling. In Fig. 5, (a) and (b) show the average IoU and recall curves with sampling rates of 0.20% and 0.25%, respectively, during the whole training process. It can be seen that both curves are spirally rising during training, and the IoU curve values are finally stable at approximately 0.85, the recall curve value eventually stabilizes at 0.94. Figure (c) shows the loss curve of the top 500 batches of iterations. It can be seen that the loss value drops rapidly during the first 100 training batches, then the change is extremely slow, and it finally stabilizes at 0.28.

B. QUALITATIVE ANALYSIS OF THE EXPERIMENTAL RESULTS

According to the Depth Fusion NMS algorithm proposed in this paper, 250 test images from the NYU Depth V2 dataset are detected in the trained fusion network and compared with the detection results using the traditional NMS algorithm, as shown in Fig. 6. The (a) rows show the detection result of the traditional NMS post-processing, and the (b) rows show the improved post-processing results based on Depth Fusion NMS.

The experiment sets the IoU threshold T to 0.6 and the depth error empirical value ε to 3. It can be seen from Fig. 6 that when there are two objects in the scene with high overlap, it is difficult to draw a box around the two objects simultaneously using the traditional NMS algorithm, but the Depth Fusion NMS algorithm can distinguish two adjacent objects with different depths. The result means that the improved NMS algorithm proposed in this paper can effectively increase the recall rate of the detection model and improve the localization accuracy of the system.

C. QUANTITATIVE COMPARISON OF THE NMS ALGORITHMS

To further verify the performance of the Depth Fusion NMS algorithm, we compare it with three algorithms on the NYU Depth V2 dataset: Greedy-NMS, Soft NMS-L [21] and Soft NMS-G [21]. In addition, we compare the performance of the four algorithms in the RGB, depth and RGB-D networks. The IoU threshold T is set to 0.6, and the parameter σ in the Soft NMS-G algorithm is set to 0.3. We calculate the average precision (AP/%) and the mean average precision (mAP/%) of the fourteen kinds categories, we also compared the average time of different algorithms with different networks, and the results are shown in Table 3.

As seen from Table 3, the Depth Fusion NMS algorithm achieves the highest AP in most categories of detection results, and the mAP is 0.8%, 0.5%, and 0.3% higher than those of the Greedy-NMS, Soft NMS-L, and Soft NMS-G algorithms, respectively. In addition, the RGB-D network is significantly more accurate for object detection than the individual RGB and depth networks. And the average detection time for one image with Depth Fusion NMS algorithm is 0.436s. Because the improved NMS algorithm mainly aims to increase the recall rate of objects with high overlap and has little effect on distant objects, so the overall performance improvement of the detection model does not seem obvious. To see the recall rate of object detection by the Depth Fusion NMS algorithm more intuitively, we selected several sets of scenes with dense objects to compare the detection results of the four NMS algorithms. As shown in Fig. 7, (a), (b), (c), and (d) are the results of Greedy-NMS, Soft NMS-L, Soft NMS-G and Depth Fusion NMS, respectively. Among them,

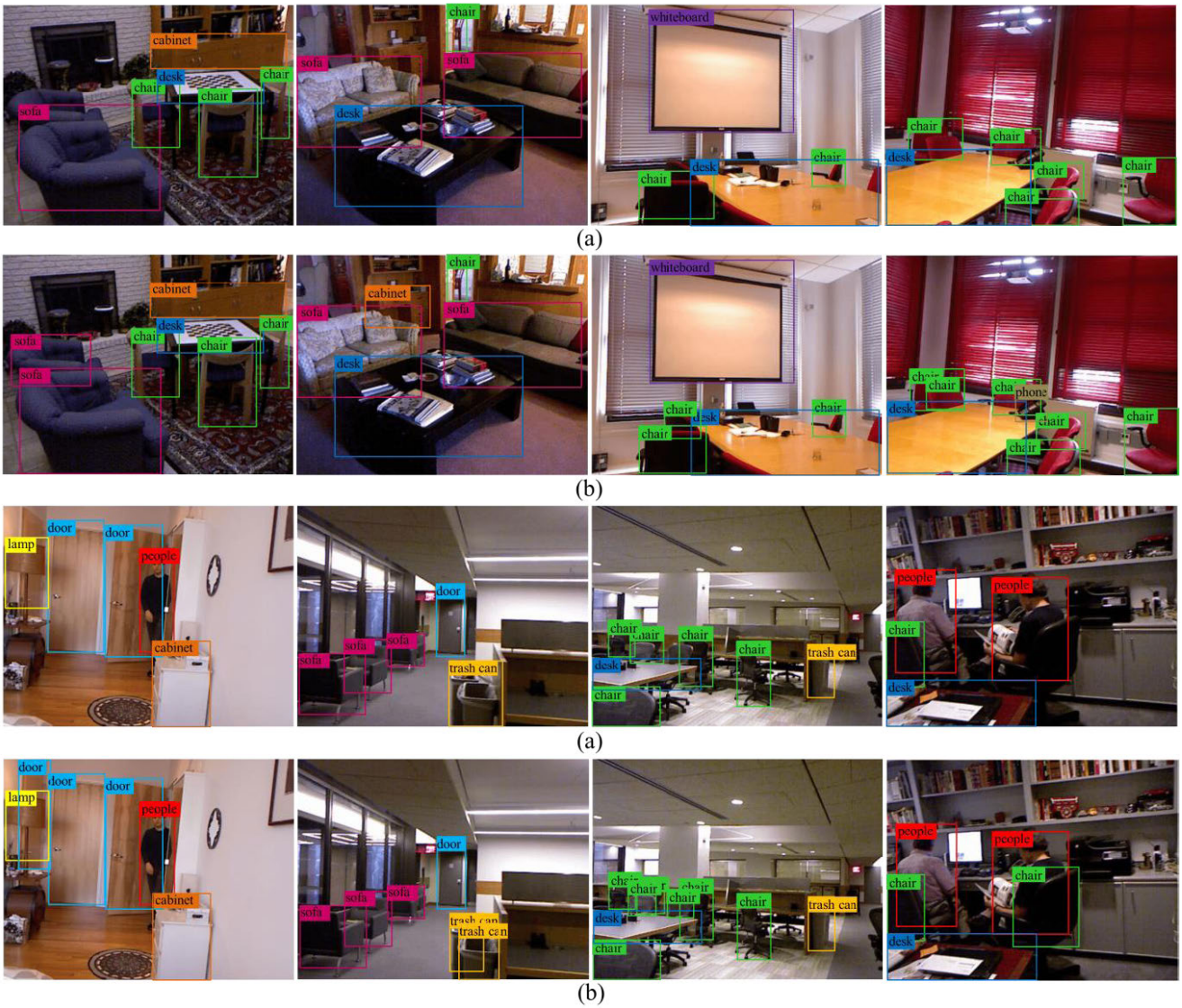


FIGURE 6. Some of the RGB-D object detection results based on the NYU Depth V2 dataset for (a), the traditional NMS algorithm and (b), the Depth Fusion NMS algorithm.

our method has an obvious detection effect on dense objects with high overlap (such as chairs and desks), which not only reduces the missing detection rate but also achieves more accurate object localization by combining with the average depth in the detection box.

In addition, we also tested the effect of different ways of fusing RGB-D information on the results. The Fig. 8 shows two different fusion modes. Fig. 8 (a) shows prophase fusion, in which the RGB and depth images are merged into a four-channel image in the data input stage for feature extraction; Fig. 8 (b) shows later fusion, in which the two modal features are respectively extracted from two convolutional neural networks and then fused in the final fully connected layer. The three RGB-D fusion models were tested using the Depth Fusion NMS algorithm proposed in this paper. Table 4 shows that the detection results of the mAP (%) for all the class, among them, ''metaphase fusion'' is the medium-term, level-by-level fusion strategy proposed in this paper. We found that the metaphase RGB-D fusion strategy and the Depth Fusion NMS algorithm can provide better detection performance than the other schemes.

To quantitatively evaluate the detection performance of all the methods under different IoU thresholds, we set the threshold variation range to 0.3 to 0.9. We obtained the mAP values of the different methods at each IoU by changing the threshold size and drew a line graph, as shown in Fig. 9. Overall, the IoU threshold is larger, the mAP of detection is smaller, mainly because more overlapping boxes are not filtered out. When the IoU threshold is low, the performance difference in the four NMS algorithms is small, but their difference becomes obvious as the threshold gradually increases by more than 0.6, and the falling gradient becomes larger

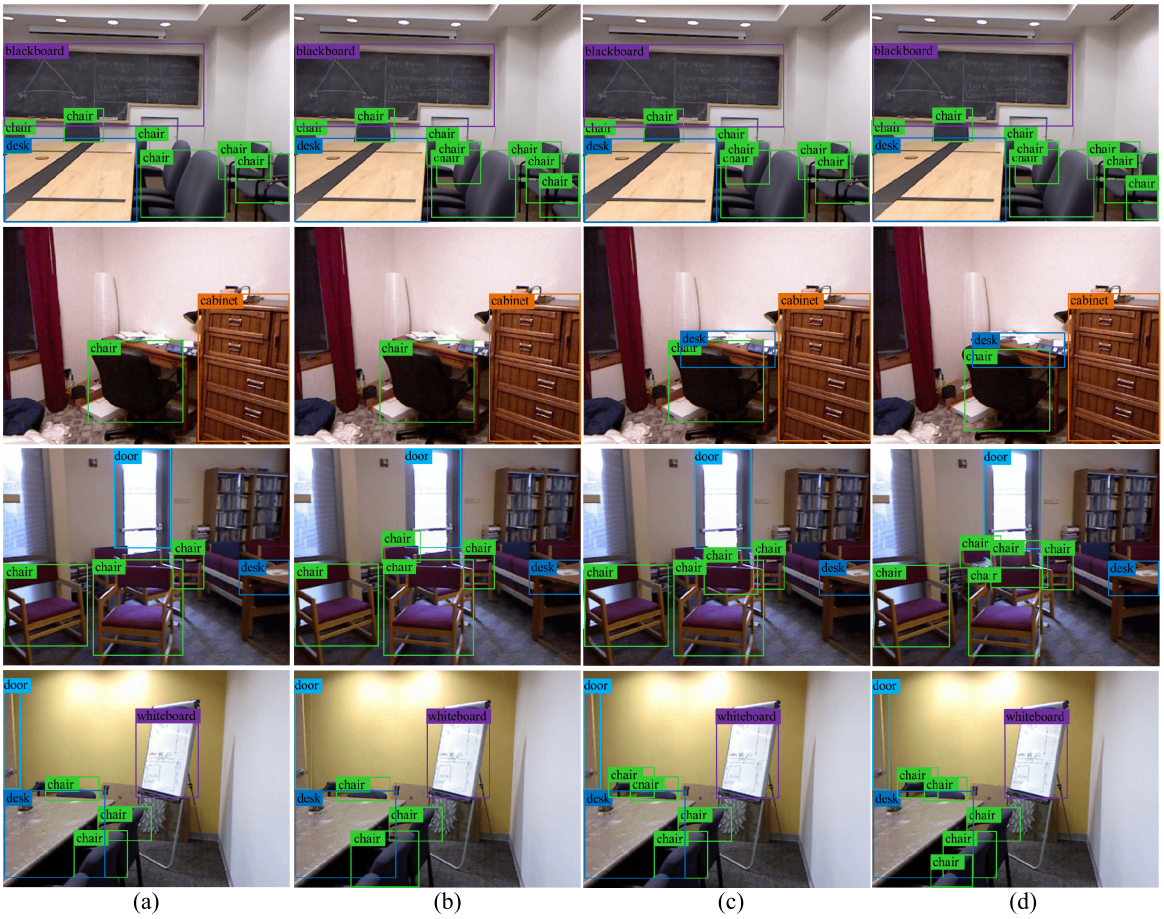


FIGURE 7. Some of the detection results from the different NMS algorithms on the NYU Depth V2 dataset. (a) the Greedy-NMS results; (b) the Soft NMS-L results; (c) the Soft NMS-G results; and (d) the Depth Fusion NMS results.

TABLE 3. AP (%) and mAP (%) for all categories achieved by the four algorithms.

Method / Category	Greedy-NMS			Soft NMS-L			Soft NMS-G			Depth Fusion NMS
	RGB	Depth	RGB-D	RGB	Depth	RGB-D	RGB	Depth	RGB-D	RGB-D
tv	73.8	62.1	77.2	73.5	63.6	**78.3**	73.8	62.7	78.1	**78.3**
chair	74.5	66.2	78.4	74.9	67.0	78.9	75.2	67.4	79.2	**80.3**
desk	72.7	62.7	76.3	72.1	62.9	76.3	73.0	62.9	76.9	**77.6**
whiteboard	59.2	50.3	61.7	59.6	50.7	61.7	59.9	50.7	**62.4**	61.9
door	66.3	58.6	70.9	66.7	59.4	71.5	66.7	59.1	71.5	**72.4**
trash can	63.4	56.0	67.1	63.0	55.6	66.5	63.8	56.3	**67.7**	**67.7**
people	75.1	66.7	78.6	76.0	67.5	**79.8**	75.5	67.0	79.3	79.3
blackboard	57.1	50.3	60.5	57.4	50.5	60.5	57.1	50.5	60.5	**61.1**
cabinet	68.2	57.9	71.6	68.8	57.9	71.9	68.5	58.3	**72.2**	**72.2**
lamp	64.8	56.7	68.0	64.5	56.7	68.0	65.3	57.0	**68.9**	68.5
sofa	72.3	61.5	74.9	72.3	61.8	74.9	72.1	61.5	74.3	**75.6**
bed	70.9	61.8	73.4	70.9	62.1	73.4	70.9	62.1	73.4	73.4
phone	56.7	44.2	58.5	57.0	44.6	**59.3**	56.7	44.2	58.9	58.9
toilet	57.6	51.0	61.7	57.1	50.8	61.7	57.8	51.3	62.0	**62.2**
mAP	66.6	57.6	69.9	66.7	58.0	70.2	66.9	58.0	70.4	**70.7**
average time(s)	0.041	0.039	**0.367**	0.044	0.040	0.512	0.042	0.041	0.479	0.436

when the IoU threshold exceeds 0.7. The results show that the Depth Fusion NMS algorithm proposed in this paper has a better object detection performance under a larger IoU threshold.

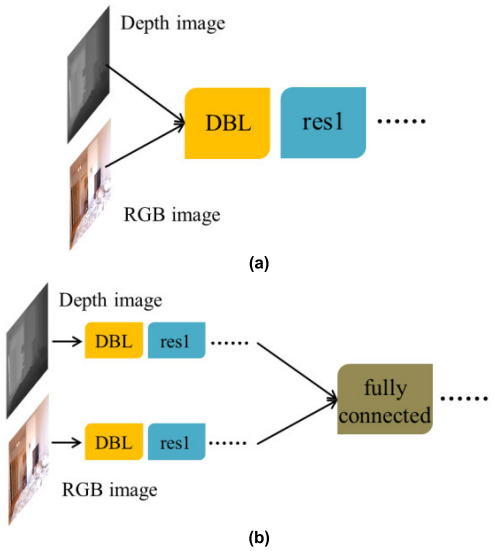


FIGURE 8. Schematic diagram of two RGB-D fusion models: (a) prophase fusion and (b) later fusion.

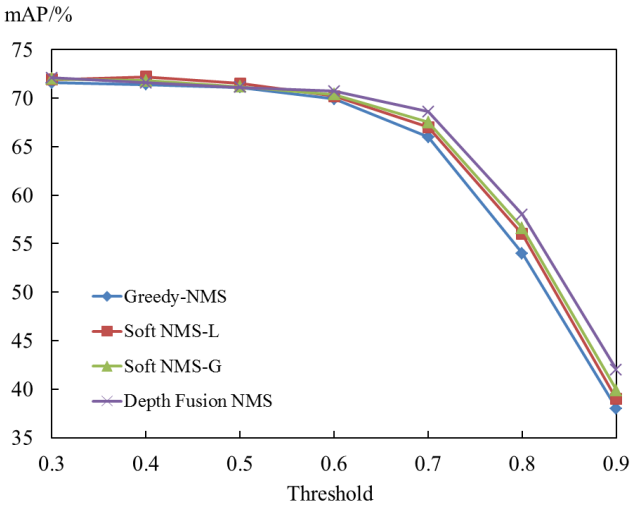


FIGURE 9. mAP of all the NMS algorithms under different IoU thresholds.

TABLE 4. mAP(%) of the different fusion strategies for the four NMS algorithms.

	Greedy-NMS	Soft NMS-L	Soft NMS-G	Depth Fusion NMS
Prophase fusion	67.1	67.8	67.3	68.1
Later fusion	67.7	68.2	69.0	68.9
Metaphase fusion	69.9	70.2	70.4	70.7

V. CONCLUSION

The post-processing stage is an indispensable step in the current popular object detection method. As a classic post-processing method, NMS has the problems of insufficiently eliminating missed and false detections due to the single constraint condition and improper IoU threshold selection. In this paper, based on the advantages of depth images in RGB-D object detection, we designed an improved NMS algorithm that depends on depth fusion, which increases the discrimination condition of objects based on the depth information. The experimental results based on the NYU Depth V2 dataset show that compared with Greedy-NMS, Soft NMS-L and Soft NMS-G, the proposed algorithm a significantly improves the detection of dense objects with high overlap at higher IoU thresholds. It can effectively reduce the object missing and false detection rate, thereby improving the accuracy of the RGB-D object detection model.

Like the traditional non-maximum suppression algorithm, the Depth Fusion NMS algorithm also faces the problem of IoU threshold selection, and it is difficult to avoid the missed detection of highly overlapping objects with similar depths. Therefore, we will continue to research how to simplify the IoU threshold-setting process of the NMS algorithm and the missing detection of near-depth objects.

REFERENCES

[1] H. Zhang, K. Wang, and F. Wang, "Advances and perspectives on applications of deep learning in visual object detection," *Acta Automatica Sinica*, vol. 43, no. 8, pp. 1289–1305, 2017.
[2] W. Q. Zhao and H. Yan, "Penalty non-maximum suppression in object detection," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, Guangzhou, China, 2018, pp. 90–102.
[3] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Hong Kong, Aug. 2006, pp. 850–855.
[4] B. Ma, Z. Liu, F. Jiang, Y. Yan, J. Yuan, and S. Bu, "Vehicle detection in aerial images using rotation-invariant cascaded forest," *IEEE Access*, no. 7, pp. 59613–59623, 2019.
[5] J. Chen and X. Ye, "Improvement of non-maximum suppression in pedestrian detection," *Nat. Sci.*, vol. 41, no. 3, pp. 371–378, Mar. 2015.
[6] Z. Wang, J. Lu, R. Lin, J. Feng, and J. Zhou, "Correlated and individual multi-modal deep learning for RGB-D object recognition," 2016, *arXiv:1604.01655*. [Online]. Available: https://arxiv.org/abs/1604.01655
[7] X. Xu, Y. Li, G. Wu, and J. Luo, "Multi-modal deep feature learning for RGB-D object detection," *Pattern Recognit.*, vol. 72, pp. 300–313, Dec. 2017.
[8] L. Schneider, M. Jasch, B. Fröhlich, T. Weber, U. Franke, M. Pollefeys, and M. Rätsch, "Multimodal neural networks: RGB-D for semantic segmentation and object detection," in *Proc. Scand. Conf. Image Anal.* Cham, Switzerland: Springer, 2017, pp. 98–109.
[9] M. R. Loghmani, M. Planamente, B. Caputo, and M. Vincze, "Recurrent convolutional fusion for RGB-D object recognition," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2878–2885, Jul. 2019.
[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767
[11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, vol. 2012, pp. 746–760.
[12] Q. Zhang, C. B. Zhang, and Z. H. Chen, "A simplified non-maximum suppression with improved constraints," *J. Univ. Sci. Technol. China*, vol. 46, no. 1, pp. 6–11, Jan. 2016.
[13] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based CNN for ship detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 900–904.
[14] S. Qiu, G. Wen, Z. Deng, J. Liu, and Y. Fan, "Accurate non-maximum suppression for object detection in high-resolution remote sensing images," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 238–247, Sep. 2018.
[15] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.
[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2015, pp. 91–99.

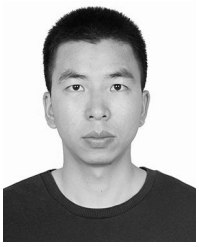
[18] Y. He, X. Zhang, M. Savvides, and K. Kitani, "Softer-NMS: Rethinking bounding box regression for accurate object detection," 2018, *arXiv:1809.08545v1*. [Online]. Available: https://arxiv.org/abs/1809.08545v1?source=post_page
[19] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Cham, Switzerland: Springer, Nov. 2016, pp. 213–228.
[20] G. Huang, Z. Liu, V. D. Laurens, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[21] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5561–5569.

DECHENG WANG was born in Baiyin, China, in 1994. He received the B.S. degree in measurement and control engineering from Space Engineering University, Beijing, China, in 2017, where he is currently pursuing the M.S. degree with the Academy of Space Information.

His research interests include RGB-D object detection and computer vision.

XIANGNING CHEN was born in Chongren, China, in 1962. He received the B.S. and M.S. degrees from The Academy of Geomatics, China, in 1984 and 1990, respectively, and the Ph.D. degree from Information Engineering University, China, in 1998.

In 2004, he joined Space Engineering University, where he involved extensively in research works in the fields of computer vision and remote sensing. He is currently a Professor with the Academy of Space Information. His research interests include optical remote sensing, image processing, and computer vision. He is a member of the Chinese Society for Geodesy, Photogrammetry and Cartography (CSGPC) and the National Remote Sensing Center of China (NRSC).



HUI YI was born in Yichun, China, in 1991. He received the B.S. degree in communication engineering from Nanchang University, Nanchang, China, in 2009, and the M.S. degree in communication and information system from the Space Engineering University, Beijing, in 2013, where he is currently pursuing the Ph.D. degree with the Academy of Space Information.

His research interests include 3D reconstruction and computer vision.

FENG ZHAO was born in Henan, China, in 1989. He received the B.S. degree in microelectronic Technology from Tianjin University, Tianjin, China, in 2011, and the M.S. degree in geomatics engineering from Information Engineering University, in 2014. He is currently pursuing the Ph.D. degree with the Academy of Space Information, University of Space Engineering, Beijing.

His research interests include photoelectron integration and information processing.