Multi-scale Feature Extraction and Fusion Net: Research on UAVs Image Semantic Segmentation Technology

Xiaogang Li[1], Di Su[2], Dongxu Chang[2], Jiajia Liu[1], Liwei Wang[1], Zhansheng Tian[1], Shuxuan Wang[3] and Wei Sun[3,*]

[1]*Henan Jiuyu Enpai Electric Power Technology Co., Ltd., 450000, Zhengzhou, China*
[2]*State Grid Henan Electric Power Company, 450000, Zhengzhou, China*
[3]*Xidian University, School of Aerospace Science and Technology,710071, Xi'an, China*
E-mail: wsun@xidian.edu.cn
[*]*Corresponding Author*

Received 20 October 2022; Accepted 22 November 2022;
Publication 14 January 2023

Abstract

Since UAV aerial images are usually captured by UAVs at high altitudes with oblique viewing angles, the amount of data is large, and the spatial resolution changes greatly, so the information on small targets is easily lost during segmentation. Aiming at the above problems, this paper presents a semantic segmentation method for UAV images, which introduces a multi-scale feature extraction and fusion module based on the encoding-decoding framework. By combining multi-scale channel feature extraction and multi-scale spatial feature extraction, the network can focus more on certain feature layers and spatial regions when extracting features. Some invalid redundant features are eliminated and the segmentation results are optimized by introducing global context information to capture global information and detailed information.

Journal of ICT Standardization, Vol. 11_1, 97–116.
doi: 10.13052/jicts2245-800X.1115

Moreover, one compares the proposed method with FCN-8s, MSDNet, and U-Net network models on the large-scale multi-class UAV dataset UAVid. The experimental results indicate that the proposed method has higher performance in both MIoU and MPA, with an overall improvement of 9.2% and 8.5%, respectively, and its prediction capability is more balanced for both large-scale and small-scale targets.

Keywords: Semantic segmentation, drone image, deep learning, multi-scale feature extraction, contextual information.

1 Introduction

Before UAVs perform various tasks, they need to perceive the task environment through sensors, obtain a map of the task area, and understand target and threat information. Vision-based sensing methods are widely used due to their strong anti-interference ability, low cost, and easy deployment. Among them, the use of pixel-level semantic segmentation technology is the main way of scene cognition.

These days, convolutional neural networks are often used in image semantic segmentation tasks. After long-term research and analysis, Long [1] established a Fully Convolutional Neural (FCN) Network which can adapt to input images of arbitrary size and also improves the feature roughness problem caused by upsampling. The FCN network effectively improves the accuracy of region-based segmentation, but it also has certain limitations. After the convolution and pooling operations in the network, the size of the original image will be significantly reduced, and the low-resolution feature representation will cause the loss of image detail information, thereby reducing the segmentation accuracy [2]. Badrinarayanan [3] proposed a SegNet network model based on an encoder-decoder framework in 2015. Although the multi-layer max-pooling and downsampling operations in the SegNet network can be robust to segmentation tasks due to their translation invariance, they result in the loss of feature map size and spatial information. To improve the above problems, the FCN algorithm-based PSPNet proposed by Zhao [4] uses global average pooling operation (GAP) and feature fusion operation to integrate the contextual information from different regions and model the global contextual information.

The Deeplab series proposed by the Google team continuously improves segmentation accuracy. The Deeplabv1 network [5] designs an atrous

convolution to exponentially expand the receptive field of the network without losing resolution and raising the computational burden. The Deeplabv2 [6] network proposes atrous spatial pyramid pooling (ASPP) in the spatial dimension. ASPP consists of atrous convolutions with different dilation rates to form a multi-scale processing module, resulting in more accurate segmentation results. The Deeplabv3 [7] network improves the ASPP module to combine four atrous convolutions with different sampling rates in a cascaded and parallel manner to encode contextual information at different scales. Subsequently, based on Deeplabv3, the Deeplabv3+ [8] network proposed in 2018 added a simple and effective decoding module to fine-tune the segmentation results, especially in the boundary part of the segmented object, the segmentation effect was significantly improved. In addition, Deeplabv3+ further uses the Xception model and Depthwise Separable Convolution and combines ASPP and a simple decoding module to obtain a faster and stronger encoding-decoding network framework, but the amount of computation also increases. big. Other studies [9–12] use an encoding-decoding structure, perform downsampling during the encoding process, gradually reduce the resolution of the feature map, continuously upsample during the decoding process, gradually restore the image size, and finally achieve high-resolution Semantic segmentation.

However, after an in-depth study of the currently used semantic segmentation methods, it was found that there are still many difficulties in the field of segmentation. Since segmentation scenes are usually complex and diverse, commonly used semantic segmentation methods cannot achieve high accuracy in every scene [13]. For example, the existing segmentation methods are prone to losing small-scale information for UAV aerial images thus making it difficult to earn accurate segmentation results. Meanwhile, the difference in target scale in the images captured by the UAV is hundreds of times different, and it is often difficult for small and weak targets to retain effective features.

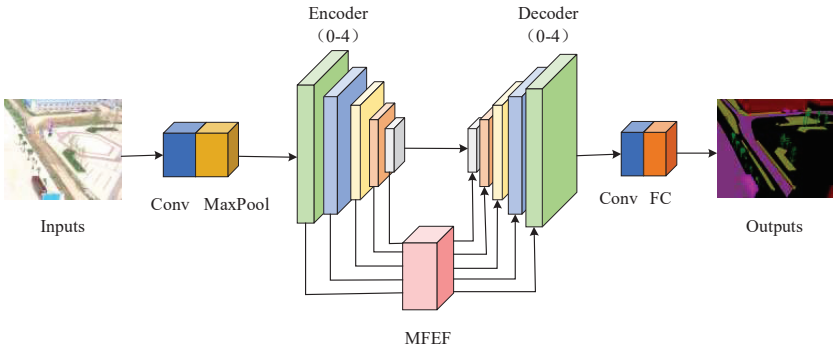
In this article, one presents a semantic segmentation method of UAV images based on multi-scale feature extraction and fusion (MFEF), which combines multi-scale channel feature extraction and multi-scale spatial feature extraction to effectively fuse detailed information and semantic information. The problem of poor segmentation performance of the shallow-level algorithm model can better restore the segmentation details of more targets, and effectively solve the problem of large changes in the scale and resolution of UAV aerial images.

2 Multi-scale Feature Extraction and Fusion Net Algorithm Design

2.1 The Overall Structure of the Network Model

There are still many deficiencies in the existing image segmentation methods for the segmentation of UAV aerial image data. First, for high-resolution images of oblique viewing angles captured by drones, the size of objects at different distances may vary significantly. Large-scale variations of objects in UAV aerial images have an impact on the precision of predictions. In the network, each output pixel in the final prediction layer has a fixed receptive field, formed by pixels in the original image that may have an impact on the final prediction of that output pixel. When the objects are too small, the neural network may learn noise from the background. When the objects are too large, the model may not get enough information to correctly infer the labels.

Motivated by these problems, this section presents a semantic segmentation algorithm for UAV images based on MFEF. The overall structure of the network model is demonstrated in Figure 1, which is an encoder-decoder framework, and skip connections are used to transfer the information between the encoding layer and the decoding layer. Among them, the encoder continuously downsamples the features to obtain the semantic features of the appropriate scale target. In addition, the decoder can also continuously upsample the features through a multi-scale feature extraction fusion module, thus gradually recovering the image resolution.



The above network model introduces a MFEF module, which uses multi-scale channel feature extraction and multi-scale spatial feature extraction to synergistically optimize the extracted complementary information, thereby learning more detailed feature representations. Specifically, the input image

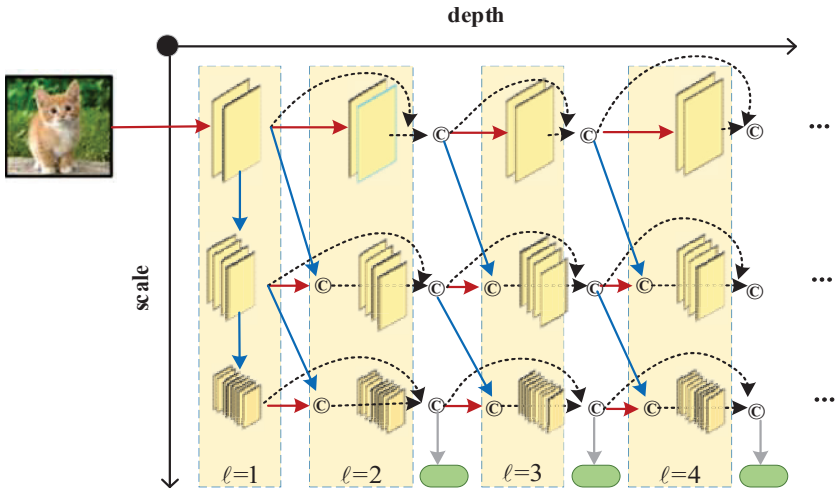


Figure 1 MFEF Net and MSDNet structure.

is handled through the convolution layer and the max pooling layer and then enters the encoder, which performs feature extraction through a down-sampling operation. Then, a MFEF module is added before the decoder of each upsampling layer, and the processed image features are sent to the decoding layer, and the decoder gradually restores the image resolution through successive upsampling layers. During this process, the image data is processed by batch normalization and ReLU activation function each time it passes through the convolutional layer. Each upsampling layer doubles the feature size and reduces the number of channels by half. In addition, through the fusion feature output by the MFEF module, the encoding results with the same feature size are fused in the form of a skip connection. This paper uses the sum operation of the corresponding position elements to realize the skip connection.

2.2 The Specific Module Network Structure

In the following, we present a MFEF module by combining multi-scale channel feature extraction and multi-scale spatial features, which effectively fuses low-level features with more detailed information and high-level features with semantic information. The context information is introduced to enhance the poor segmentation performance of the shallow algorithm model, and to better restore more target segmentation details. The specific module network structure is given in Figure 2.

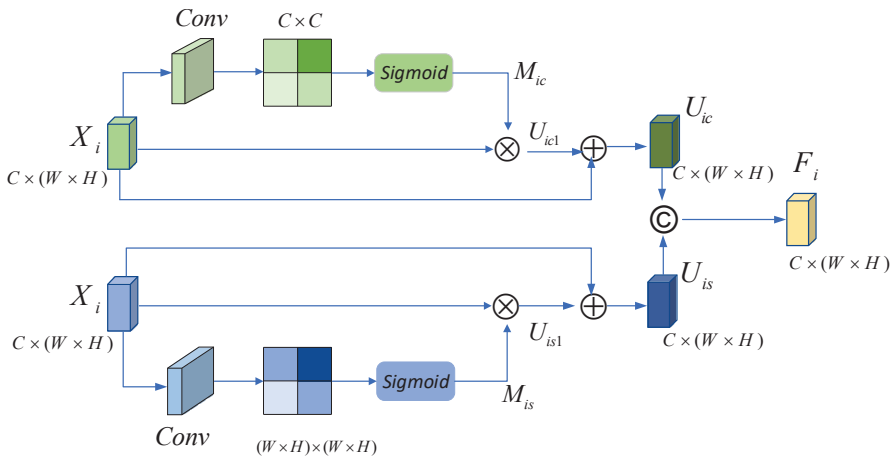


Figure 2 MFEF module.

The MFEF module is mainly divided into a multi-scale channel feature extraction module and a multi-scale spatial feature extraction module. Where $X_i \in R^{C\times H\times W}$ represents the feature from the i encoder stage, and the channel number and size of the feature are C and $H \times W$, respectively; the symbol $\otimes$ and $\oplus$ represent the multiplication and addition operations of the corresponding position element, respectively; X_i will be used as the input of the MFEF module. Among them, the first part is the multi-scale channel feature extraction module. Firstly, the major channel information of the input feature is extracted using the global average pooling method, and then it is input into a $1 \times 1 \times 1$ convolution to compress the convolution parameters, and then use the sigmoid activation function to normalize the input value to obtain a multi-scale channel feature mask M_{ic}. Then, the processed result is multiplied by the original eigenvalue to obtain an information-calibrated feature map U_{ic1}. The second part is the multi-scale spatial feature extraction module, which mainly obtains the spatial feature map by using the spatial correlation between the features on the input feature map. Compared with multi-scale channel feature extraction, multi-scale spatial feature extraction is more concerned with the spatial information of feature images. The input feature image is sequentially input into the convolutional layer and the activation layer, and output from the sigmoid layer to obtain the multi-scale spatial feature mask M_{is} after activation. Combining it with the original eigenvalues, a feature map U_{is1} can be obtained, which completes the calibration process of spatial information.

Here, to reduce the effect of background information, overcome the problem of lack of global information in the algorithm due to the characteristics of shallow layers, and improve the segmentation function of the model, context information can be added to the decoding process, and the features of each stage in the model processing process can be further processed. In the above two feature extraction, the input information X_i is the original feature block, the weighted feature maps U_{ic1} and U_{is1} are balanced, and the feature weights are re-adjusted to obtain the matching weight map U_{ic} and U_{is}.

$$U_{ic} = X_i \oplus U_{ic1} = X_i + M_{ic} \otimes X_i \tag{1}$$

$$U_{is} = X_i \oplus U_{is1} = X_i + M_{is} \otimes X_i \tag{2}$$

The higher the weight ratio is, the more important it is. Therefore, this paper sets the weight value range of the elements in M_{ic} and M_{is} from 0 to 1. When the value of a certain position is close to 1, the value of the output feature U_i increases at that position is large, and when the value of a certain position is close to 0, the output features U_i are nearly the same as the initial features X_i. In this way, after the multi-scale spatial features are processed by the fusion module, they are adapted to the multi-scale channel features, and the information of the original features is also retained, which is more conducive to network learning and presents better feature effects.

Further, the output features $F_i \in R^{C\times H\times W}$ of the fusion module extracted from multi-scale features can be obtained by the following algorithm:

$$F_i = f^{1\times 1}(concat(U_{ic}, U_{is})) \tag{3}$$

Among them, $concat$ represents the splicing action performed in the channel, $f^{1\times 1}$ represents the nonlinear transformation effect, including the ReLU activation layer, the convolutional network with stride set to 1, and the batch normalization operation. After this processing, the feature size remains unchanged, but the total number of channels changes for the initial half.

The global features and detailed features are reflected in the whole process, which greatly reduces the calculation amount of the proposed model and enhances the segmentation capacity of the model with less computation through the connection of the upper and lower layers of information. At the same time, adding the initial feature maps of each stage is more conducive to enhancing the global feature extraction ability, and the MFEF module further refines the output fusion features, thereby enhancing the interaction between data. It guides the analysis and prediction of features, and ultimately improves the segmentation characteristics.

3 Semantic Segmentation Experiment

3.1 Experiment Preparation

All semantic segmentation experiments performed in this paper are trained, tested, and evaluated on the lab's server for deep learning. As shown in Table 1, the environmental configuration required to conduct this experiment is listed in the table.

This paper evaluates the proposed method using the UAV semantic segmentation dataset UAVid [14]. The UAVid dataset is established for UAV semantic segmentation in complex urban scenes, focusing on 8 object categories. The datasets are captured by drones with oblique views, providing multiple representations of objects with rich scene backgrounds, with large variations in spatial resolution. The dataset has 420 high-quaity 4K images (4096×2160 or 3840×2160) divided into training set, validation set and test set with 200, 70 and 150 images, respectively, containing the most common and representative 8 object categories, namely buildings, roads, trees, low vegetation, static vehicles, moving vehicles, people and sundries. Examples of different classes are given in Figure 3. The definition description of each class is given in Table 2.

All models in this experiment are performed on the PyTorch platform with i7-13700@5.2GHz, four NVIDIA GTX 1080Ti GPUs. The network model given in this chapter uses a small batch stochastic gradient descent method with a batch of 8, a momentum of 0.9, and a weight decay of 0.0001 to optimize the model. The data preprocessing process uses data augmentation strategies, such as random flipping, random cropping, etc. The initial size of the UAVid dataset is usually 4096×2160 or 3840×2160. During training, in order to facilitate batch processing, the images of the dataset are randomly cropped and then scaled to a resolution of 1024×1024. As the input of the

Table 1 Lab environment

Name	Configuration
Operating system	Ubuntu18.04
GPU model	NVIDIA GTX1080Ti
Video memory size	11G
Python version	Python3.6
CUDA version	CUDA10.0
cuDNN version	cuDNN7.6
PyTorch version	PyTorch1.10.1

Figure 3 Experimental dataset.

Table 2 Experimental dataset class definitions

Category	Description
Building	Homes, garages, skyscrapers, security booths and buildings under construction. Freestanding walls and fences are not included.
Road	A road or bridge surface on which a car can legally travel. Parking not included.
Tree	Tall trees with crowns and trunks.
Low Vegetation	Grasses, Shrubs and Shrubs.
Static Car	Immobile vehicles, stationary buses, trucks, trains, cars and tractors are included, but bicycles and motorcycles are excluded.
Moving Car	Moving cars, including moving buses, trucks, cars and tractors. Bicycles and motorcycles are not included.
Humans	Pedestrians, cyclists and all other people engaged in different activities.
Background Clutter	All targets that fall into any of the categories above.

algorithm, overlapping pixels are 24×24. Take the initial learning efficiency to 0.001.

In order to test the semantic segmentation performance of the network model, this paper takes Mean Pixel Accuracy (MPA) and Mean Intersection Over Union (MIoU) as the evaluation criteria.

The IoU and PA represented using the confusion matrix are shown in Figure 4. TP(True Positive) represents the true example, which means that the target includes both the actual and the expected. TN stands for True Negative, which means that the goal includes the actual situation but not the expectation. FP (False Positive) means false positives, meaning that the target

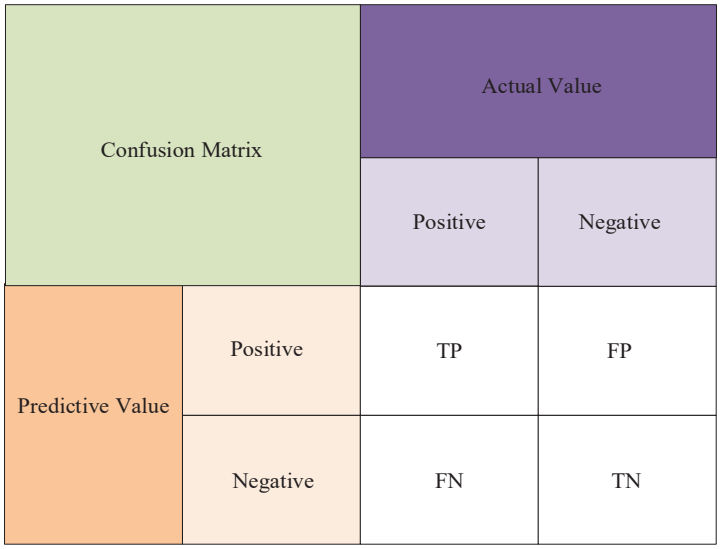


Figure 4 Confusion matrix.

contains expectations but not actuals. FN (False Negative) represents a false negative, meaning that the target contains neither actual nor expected.

Intersection and Union Ratio (IoU): The intersection and sum operation are performed on two sets containing the actual value and the expected value, and then the ratio of the two is calculated. The calculation formula of the intersection and union ratio is shown in formula (4):

$$IoU = \frac{TP}{TP + FP + FN} \tag{4}$$

Mean intersection-over-union (MIoU): Calculate the IOU value on each category, and then average the IOU for all categories. Its calculation formula is as follows (5):

$$MIoU = \frac{1}{k} \sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}} \tag{5}$$

Pixel Accuracy (PA): Calculates the proportion of correctly predicted pixels out of total pixels. Its calculation formula is as follows (6):

$$PA = \frac{TP + TN}{TP + TN + FT + FN} \tag{6}$$

Mean Pixel Precision (MPA): Calculate the PA value on each class, and then average the PA across all classes. Its calculation formula is as follows (7):

$$MPA = \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}} \tag{7}$$

In Equations (5) and (7), k represents the number of categories, j stands for the number that belongs to category but is predicted to be category, i represents the real number, and p_{ij} represents false positives and false negatives.

3.2 Analysis of Experimental Results

The FCN-8s, MSDNet and U-Net algorithms are selected as the comparators with the proposed algorithm for MIoU and MPA on the UAVid dataset. The comparison algorithms used are all introduced in the previous chapter. The results are given in Tables 3–4.

From Table 3, among all the compared models, our method has the best performance in terms of MIoU, with an overall improvement of 9.2%. For each category evaluation, our method ranks first in the categories of buildings, static vehicles, vegetation, humans, and dynamic vehicles, with improvements of 6.2%, 17.7%, 0.5%, 15%, and 20%, respectively. The most obvious improvement is for dynamic vehicles, while significant improvements are also achieved in the categories of humans, static vehicles, and buildings. The average pixel accuracy of different algorithms is shown in Table 4.

Table 3 MIoU comparison result (unit: %)

Method	Clutter	Building	Road	Static Car	Tree	Vegetable	Human	Moving Car	Mean
FCN-8s	47.4	45.0	**30.0**	10.0	**59.1**	48.3	11.0	9.0	32.5
MSDNet	37.1	47.2	20.0	30.0	47.6	44.9	20.0	10.0	32.1
U-Net	**49.5**	50.8	21.0	10.0	58.0	32.8	20.0	10.0	31.5
Ours	43.7	**57.0**	26.3	**47.7**	44.7	**48.8**	**35.0**	**30.0**	**41.7**

Table 4 MPA comparison result (unit: %)

Method	Clutter	Building	Road	Static Car	Tree	Vegetable	Human	Moving Car	Mean
FCN-8s	52.2	47.1	**50.0**	20.0	**63.9**	56.4	29.0	11.0	41.2
MSDNet	44.2	41.6	24.0	26.0	57.4	53.2	30.0	30.0	38.3
U-Net	**56.9**	55.9	30.0	10.0	67.2	49.5	30.0	20.0	39.9
Ours	55.3	**62.1**	31.8	**48.6**	52.3	**76.1**	**36.5**	**35.0**	**49.7**

From the results in Table 4, among all the compared models, the method proposed performs the best in terms of MPA, with an overall improvement of 8.5%. For each category evaluation, our method ranks first in the categories of buildings, static vehicles, vegetation, humans, and dynamic vehicles, with improvements of 6.2%, 22.6%, 19.7%, 6.5%, and 5.0%, respectively. The best enhancement is achieved in static vehicles, while significant improvements are also achieved in the categories of buildings, vegetation, humans, and dynamic vehicles.

Overall, the method proposed in this article has a more balanced prediction capability for large-scale and small-scale targets. By merging global features and detail features, our method achieves good prediction results in the classification of moving and static vehicles. For the human category, the segmentation performance of our method is higher than the other three networks, which reflects the superiority of the MFEF module in dealing with small-scale objects. At the same time, good prediction results were also obtained in buildings and vegetation.

One compares the segmentation results of the developed method with those of FCN-8s, MSDNet and U-Net on the UAVid dataset, and the visualization results are given in Figures 5–8.

Figure 5 shows the predicted segmentation performance of each network model for buildings and roads. Compared with the method in this paper, FCN8s, U-Net and MSDNet will misjudge the open space in the lower-left corner of the picture as a building or a road, while the proposed method has the smallest error and good prediction effect.

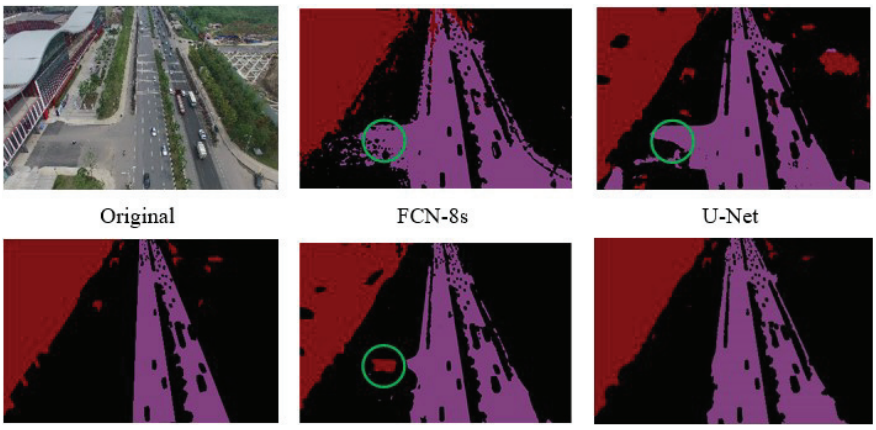


Figure 5 Prediction of buildings and roads using different networks.

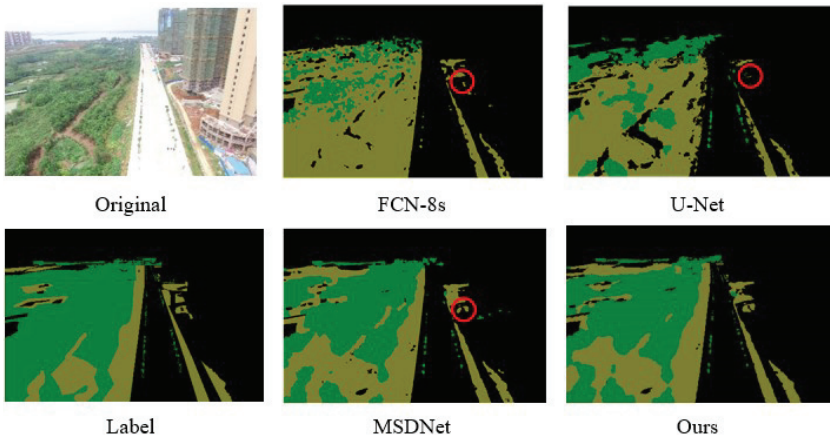


Figure 6 Prediction of trees and vegetation using different networks.

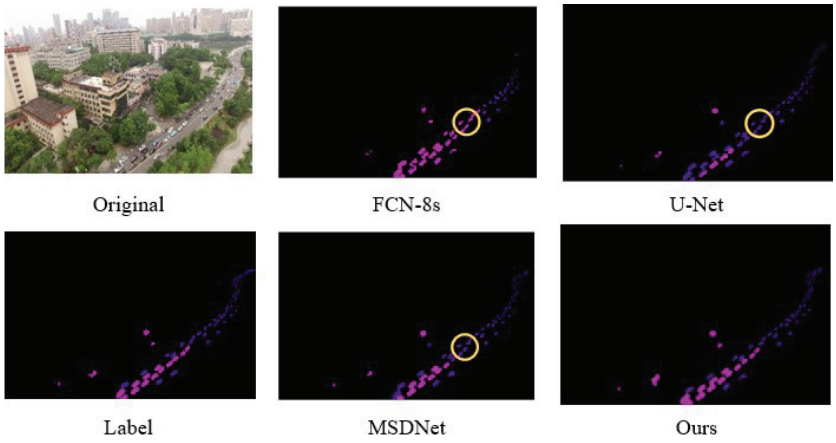


Figure 7 Prediction of moving cars and stationary cars using different networks.

Figure 6 shows the predicted segmentation performance of each network model for trees and vegetation. Compared with the method in this paper, FCN8s, U-Net and MSDNet will misjudge the vegetation in the upper right corner of the picture as the background. The proposed method is effective in the segmentation results of trees and vegetation and more accurate in segmenting the boundaries between trees and vegetation.

Figure 7 illustrates the effectiveness of different models for predicting segmentation of moving and stationary vehicles. Compared to our method, FCN-8s, U-Net, and MSDNet are less accurate in separating the boundaries

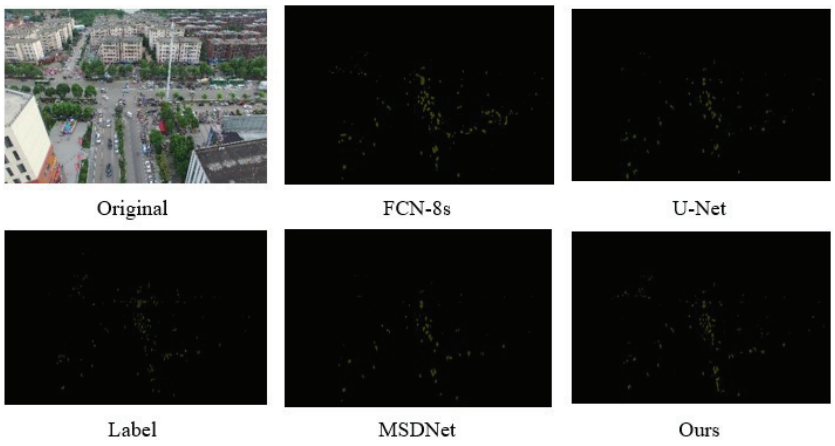


Figure 8 Prediction of humans using different networks.

of moving and stationary vehicles, erroneously identifying stationary vehicles as moving vehicles. Our method can completely separate moving vehicles from stationary vehicles and can accurately identify moving vehicles in the distance.

Figure 8 shows the predicted segmentation results for people by each network model. Due to the lack of deep extraction of original scale information and the loss of effective information of small-scale objects, FCN-8s, U-Net and MSDNet have poor segmentation results. Our method can effectively detect and segment majority of people in the image when dealing with small-scale targets, even if the targets are small, obtaining segmentation accuracy better than that of other networks.

4 Conclusion

Based on deep learning, this work applies the scene semantic segmentation technique to UAV images and achieves the following main contributions.

- One presents a new UAV image semantic segmentation algorithm. It adds a MFEF module on the basis of the encoding-decoding mode and uses the weighted allocation method to obtain different weight values of the feature map, which well balances the segmentation results of different scale targets.
- The simulation results indicate that the MFEF network achieves the best results in MIoU and MPA, which are improved by 9.2% and 8.5%,

respectively. At the same time, it also achieved the best segmentation results in the five categories of buildings, static vehicles, vegetation, humans and dynamic vehicles. The prediction results for small-scale humans increased by 15% in IoU and 6.5% in PA. %.

- According to the experimental results, the MFEF Net algorithm has the characteristics of high segmentation accuracy and precision, and can effectively handle the difficulty of large changes in the scale and resolution of UAV aerial images, which proves the rationality and feasibility of the algorithm.

Although some progress has been made, this research still has shortcomings and room for improvement, such as poor real-time performance. Therefore, the next research direction is to seek more efficient and fast segmentation in complex UAV application scenarios.

Acknowledgements

This paper was supported by the National Natural Science Foundation of China61671356: Shaanxi Provincial Key R&D Plan (2020GY-136Shaanxi Key R&D Plan Key Industry Innovation Chain Project (2019ZDLGY14-02-03,2022ZDLGY03-01) China College Innovation Fund of Production, Education and Research (2021ZYAO8004); Xi'an Science and Technology Plan Project (2022JH-RGZN-0039); Graduate Innovation Fund in Xidian University (YJS2217).

References

[1] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4): 640–651.

[2] Duan Lijuan, Sun Qichao, Qiao Yuanhua, et al. Semantic Segmentation Algorithm of RGB-D Indoor Image Based on Attention Perception and Semantic Perception [J]. Chinese Journal of Computers, 2021.

[3] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495.

[4] Zhao H S, Shi J P, Qi X, et al. Pyramid Scene Parsing Network [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawail. USA. 2017: 6230–6239.
[5] Chen L C, Papandreou G, Kokkinos I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs [J]. Computer Science, 2014(4): 357–361.
[6] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834–848.
[7] Chen L C, Papandreou G, Schroff F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation [J]. ArXiv preprint arXiv: 1706.05587, 2017.
[8] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation [C]. Proceedings of the European conference on computer vision (ECCV), 2018: 801–818.
[9] Badrinarayanan V, Kendall A. Cipolla R. SegNet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495.
[10] Ronneberger O, Fischer P, Brox T. U-Net: Convolution Networks for Biomedical Image Segmentation [C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany, 2015: 234–241.
[11] Jegou S, Drozdzal M, Vazquez D, et al. The One Hundred layers Tiramisu: Fully Convolutional Densenets for Semantic Segmentation. arXiv preprint arXiv: 1611.09326, 2016.
[12] Lin G S, Milan A, Shen C H, Reid I. RefineNet: Multi-path Refinement Network for High-resolution Semantic Segmentation. //Proceedings of the IEEE(Conference on Computer Vision and Pattern Recognition. Hawaii, USA. 2017: 5168–5177.
[13] Wang Yanran, Chen Qingliang, Wu Junjun. A Review of Image Semantic Segmentation Methods for Complex Environments [J]. Computer Science, 2019, 46(9): 36–46.
[14] Ye L, Vosselman G, Xia G, et al. UAVid: A Semantic Segmentation Dataset for UAV Imagery [J]. 2018.

Biographies

Xiaogang Li is a master's student and senior engineer. He graduated from Power System and Automation at Tianjin University in July 2010. He joined Henan Jiuyu EPRI Electric Power Technology Co., Ltd. in January 2019. He is mainly engaged in the technical service of power equipment. He is good at research on intelligent operation inspection technology of power grids, application research on new energy technology of electric power, electrical equipment tests, and fault diagnosis research.

Di Su is a master's student and senior engineer. He graduated from Motor and Electrical Equipment of North China Electric Power University in March 2006. He joined the State Grid Zhengzhou Electric Power Company in July 2006. He is engaged in production, operation and maintenance, management, organization, and personnel work. He is good at safety production, comprehensive management, and personnel management.

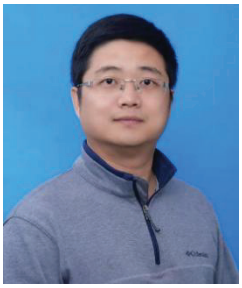
Dongxu Chang received his B.Sc. degree in Electrical Engineering from Air Force No. 1 Aviation University, Xinyang, China, in 2008. He has been engaged in UHV DC operation and maintenance for 14 years and has accumulated solid professional work experience. From Fufeng DC to Jinsu DC, and then to Tianzhong DC and Qingyu DC, he has traveled across the whole country and witnessed the growth, development, and expansion of UHV in China.

Jiajia Liu received his B.Sc. degree in automation from Zhengzhou University (Zhengzhou, China) in 2007. He is currently working as an engineer at Henan Jiuyu Enpai Electric Power Technology Co., LTD., Zhengzhou, China. He is mainly engaged in high-voltage test inspection.

Liwei Wang was born in Xuchang, China. He received his B.Sc. degree in Detection Technology and Instruments from Xidian University in 1996. He is currently the deputy manager of Henan Jiuyu Enpai Power Technology Co., Ltd. He has rich working experience in the power system-related working field, familiar with power production, production and operation, product development, and distribution network product detection. He has a solid commissioning experience with power automation equipment and has participated in the on-site commissioning of automation equipment many times. As a key member, he participated in many automation equipment R&D projects and the R&D implementation of several information projects.

Zhansheng Tian graduated from PLA Information Engineering University (Zhengzhou, China). He joined Henan Jiuyu Enpai Electric Power Technology Co., LTD. in 2010. He is currently the leader of the substation group in the equipment division of the company. He has led the diagnosis of more than 100 accidents of 110–500 kV main equipment, diagnosed many generator rotor turn-to-turn short circuit faults in the province, and participated in the preparation of the Technical Specification of Partial Discharge Online Monitoring Device for Gas Insulated Metal Enclosed Switchgear with UHF Method.

Shuxuan Wang received her B.Sc. degree from the Northeast Electric Power University, Jilin, China, in 2019, and her M.Sc. degree from the Xidian University, Xi'an, China, in 2022. She is currently working in Guangzhou Asensing Technology Co., Ltd. Her research interests include image target detection and image semantic segmentation.



Wei Sun received his B.Sc., M.Sc. and Ph.D. degrees from the Xidian University (Xi'an, China), in 2002, 2005, and 2009, respectively. He has been a professor in the School of Aerospace Science and Technology at Xidian University since 2017. His research interests include intelligent robots and unmanned aerial systems.