FAST MULTI-SCALE HRNet FOR COAL AND ROCK IMAGE SEGMENTATION

Bingfei NAN[12*]

Coal and rock image segmentation is an essential part of video surveillance systems in the coal mine underground operation scene. Driven by practical applications, a novel Fast Multi-scale HRNet for the coal and rock image segmentation is proposed, which utilizes the depthwise separable convolution to speed up the segmentation processing and uses the multi-scale feature aggregation to enhance the segmentation performance. Firstly, the down-sampling operation is conducted to reduce computational cost. Then the fast high-resolution network is designed to obtain multi-scale features, which repeats multi-scale fusion with parallel structure to enhance the multiple resolution feature representation. Finally, the multi-scale feature aggregation module is designed to build dilated convolution spatial pooling pyramid, which uses a variety of scales of dilated convolution and integrates different receptive fields to extract more effective information. The proposed fast multi-scale HRNet remains the parallel structure and has faster reasoning speed than the HRNet. Extensive experiments have been conducted, and their results have shown that the proposed fast multi-scale HRNet has higher segmentation accuracy than the HRNet and the Lite HRNet, which can meet practical application needs on mobile devices with weak computing power and small storage space.

Keywords: image segmentation; multi-scale HRNet; down-sampling; depthwise separable convolution; feature aggregation

1. Introduction

The video safety production monitoring system has been widely used in the coal mine underground operation scene. The video monitoring system uses vision to realize the intelligent perception of the operational environmental status, the automatic detection and recognition of the key equipment and target status, and the intelligent monitoring and alarm of abnormal conditions to realize the intelligent perception of the safety production process. The video surveillance system can reduce the labor intensity of the staff, and improves the remote visual intelligent control level of the working face. Among them, the semantic segmentation of the coal and rock image is the key techniques of the whole video surveillance system.

[1] Beijing Tianma Intelligent Control Technology Co., Ltd., Beijing 101300, China. e-mail: nanbf@tdmarco.com
[2] National Key Laboratory of Intelligent Coal Mining and Rock Stratum Control, Beijing 100013, China.

Image segmentation analyzes the semantic content of images or videos by using various algorithms. The traditional method are not end-to-end methods and have no speed and resource consumption advantages. The deep neural network-based methods could automatically learn image features, which are the most significant difference from the traditional image segmentation methods. This paper focuses the coal and rock image segmentation method by using the deep learning technique. At present, the deep neural network can improve the accuracy of semantic segmentation from different perspectives, such as model structure, loss function, and efficiency. However, the effectiveness of image semantic segmentation is gradually reduced in the face of various factors in complicated environments, such as unstructured, target diversification, shape irregularity, and object occlusion. The high-resolution network (HRNet) [1] performs well in many image analysis tasks. The highlight of the network structure is that it can maintain detailed representation in the training process. It connects multiple subnets and low-resolution subnets in parallel. Different from most image segmentation methods that only perform fusion once, the HRNet performs repeated multi-scale fusion by repeated information exchange and fusion. Therefore, it can perform better in most application scenarios.

Although a large number of parameters and parallel structure of HRNet greatly increase its ability to extract features from the network, the target contour information that needs to be learned from the coal and rock images does not need such strong feature extraction ability. On the contrary, the complex network structure of HRNet will limit the performance and speed of the segmentation. Therefore, a Fast Multi-scale HRNet for the coal and rock image segmentation is presented. It can significantly reduce the scale of the model while ensuring accuracy through the fast high-resolution network and the multi-scale feature aggregation module.

2. Related works

Traditional image segmentation methods usually use shallow visual features to segment different kinds of objects, and then semantic information is manually labeled for image understanding tasks. In recent years, researchers began introducing deep neural network models into semantic segmentation methods. This paper mainly focuses on the deep learning-based research works of image semantic segmentation. This kind of methods usually improve the basic network according to the specific application scenario to obtain the improvement of semantic segmentation performance. Farabet et al. made the first attempt to combine convolutional neural networks (CNN) for image segmentation [2]. They trained a multi-scale CNN using the Laplacian Pyramid [3]. They obtained the original image segmentation contour through hyperpixel and segmentation tree to supervise and

learn the multi-scale convolution network. Subsequently, more semantic segmentation network models such as DeepLab [4], PSPNet [5], and DANet [6] were proposed. Among them, the Fully Convolutional Network (FCN) was the first network using encoding and decoding methods. It has a strong ability to cope with image semantic segmentation in complicated environments. However, the accuracy of the FCN model is insufficient as it does not consider the useful contextual information.

For image semantic understanding, apart from the excellent network structure, the model is required to integrate various information, including the integration of scale space information and the balance of local and global information. Among the methods using multi-scale information, the most classic is the DeepLab series of methods proposed by Google. The DeepLab v1 [7] uses the dilated convolution and can adjust the receptive field. The DeepLab v2 [8] added a multi-view field based on the v1 version to construct a multi-scale model. This structure is known as the Atlas Spatial Pyramid Pooling (ASPP). Next, Chen et al. proposed DeepLab v3 [9] to improve the ASPP module based on the v2 version. To improve the defect of the DeepLab v3 that the details are ignored, Chen et al. proposed DeepLab v3+ [10]. The v3+ version uses the improved version of Xception as the basic network. The Experimental results have shown that the DeepLab v3+ performs better than ResNet101. In addition, a decoding module based on the DeepLab v3 version is designed to retain the edge details of objects further.

HRNet (high resolution Net) [1] firstly uses a high resolution branch and continuously adds high resolution to low resolution subnets to maintain high resolution feature map. Multiple branches in multiple stages are connected in parallel. Furthermore, this network can obtain accurate spatial semantic information and has a strong expressive ability, which basically solves the problems of insufficient precision. However, information exchange and feature fusion between different branches will also produce many parameters and redundant information. Recently, for the extensive computationof the HRNet, Gao et al. proposed Lite-HRNet (Lightweight High Resolution Network) [11]. This network applied the shuffle blocks to HRNet, and has achieved better performance than other outstanding networks. Since the pointwise 1×1 convolution has become the computing bottleneck of the network, the Lite-HRNet network introduces a lightweight computing unit to replace the expensive point-to-point convolution. The pointwise 1×1 convolution is a quadratic time complexity, and the calculation amount of general channel weighting is far less than that of point-to-point convolution. To compensate for the pointwise convolution used in exchanging information between different channels, the Lite-HRNet network adaptively calculates the weight of each channel to exchange information across channels with different resolutions. This network improves the problems of the original HRNet

network, which is computationally expensive and has large storage requirement, but the inference speed of the network has not been effectively improved.

In recent years, researchers have proposed some method for coal-rock segmentation. Xu et al. [12] presents a real-time coal-rock interface segmentation network based on the improved YOLO and the lightweight bilateral structure. The fuse attention mechanism's coal rock full-scale network (FAM-CRFSN) is proposed to improve the performance of coal-rock image segmentation [13]. Sui et al. Use the DeepLabv3+ model to perform coal-rock semantic segmentation [14]. Although the above methods have achieved good performances, but they are easily influenced by illumination conditions.

3. The proposed semantic segmentation method

The whole flowchart of the proposed coal and rock image segmentation method is illustrated in Fig. 1. It consists of the down-sampling feature extraction module, the fast high-resolution network, and the multi-scale feature aggregation module. The down-sampling feature extraction module receives the original coal and rock image and output the down-sampling feature map. The resolution of the image is reduced in this process for computation efficiency. The fast high resolution network module is responsible for obtaining the multi-scale features, where the amount of parameters of HRNet is reduced with depthwise separable convolution strategy. Finally, the effective information of multi-scale features is fused to acquire better segmentation results. Detailed descriptions of the above three modules can be found in the subsequent sections.

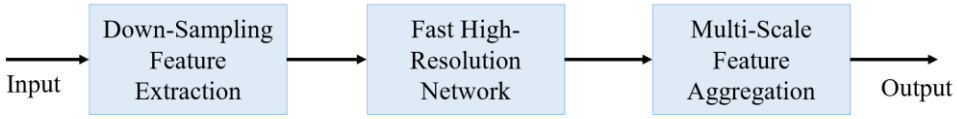


Fig. 1. The whole flowchart of the proposed FM-HRNet

3.1. The down-sampling feature extraction module

Semantic segmentation for coal and rock images generally uses relatively few pixel categories and images with high resolution is not required. It is unnecessary for a massive amount of computation to be used for redundant information from those high-resolution pixels. In this work, the down-sampling feature extraction module is applied to effectively filter redundant information to reduce the waste of computation resources.

Currently, there are normally two down-sampling approaches for convolutional neural network with similar effects: using a pooling layer or using a convolutional layer. In this work, the latter method is applied for the reason that the

multi-layer stacked convolutional layers can achieve the effect of nonlinear mapping provided by the pooling layer, while retaining more semantic information. To reduce the computational load, the original image is first sent to the down-sampling feature extraction module. The down-sampling feature extraction module is illustrated in Fig. 2. The step size of the residual convolution of the down-sampling module is 2, and the padding is 1. After several basic convolutional operations and down-sampling, the feature map's resolution is 1/8 of the original image and 64 channels is obtained.

The down-sampling feature extraction module has four stages with a serial structure, and each stage is connected in series. To enable the network to learn more abundant features, such as different directional features and different frequency features, it is necessary to increase the feature channels. So the feature channels is gradually increased and the output has 64 channels in the down-sampling feature extraction module. Finally, these down-sampling features will be input into our proposed fast high-resolution network for richer feature extraction and semantic analysis. The down-sampling feature map can reduce redundant information efficiently.

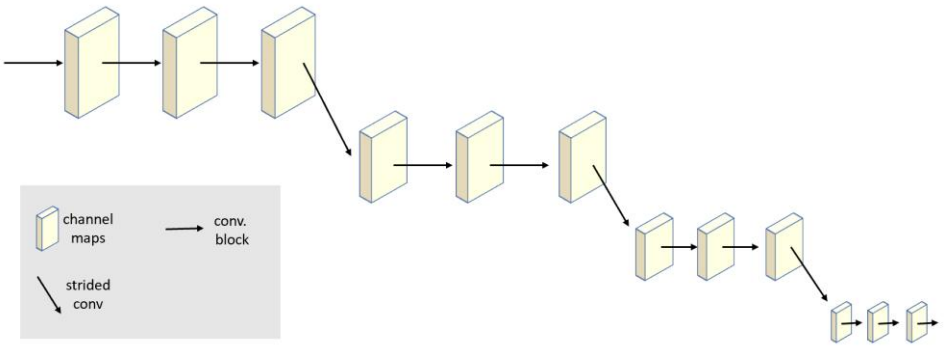


Fig. 2. The down-sampling feature extraction module.

3.2. The fast high-resolution network

In this section, the HRNet is selected as the baseline. The HRNet has a parallel network structure that connects high resolution to low resolution at the same time. This network can maintain high resolution, instead of gradually recovering feature maps from low resolution to high resolution, so that the prediction results are more accurate in space. Repeated multi-scale fusion with parallel structure will enhance the high-resolution feature representation and lead to more accurate semantic segmentation results. To further speed up the image segmentation algorithm, an optimized structure called Fast HRNet is developed. In

this new design, the utilization of depthwise separable convolution significantly accelerates the computation speed.

The fast high-resolution network consists into four stages, as shown in Fig. 3 illustrates. At each stage, a new branch for extracting features from smaller dimensions is added. Starting from the high-resolution subnetwork, the subnetworks are added to low resolution, ending with several subnetworks connected in simultaneously. Then, rich high-resolution representations are produced by repeatedly performing multi-scale fusions. The first stage and the second stage are only performed once, the third stage is repeated four times, and the fourth stage is repeated three times. To learn deeper image feature and enlarge the receptive field, a 2 step 3×3 convolution is performed for deeper downsampling. 1/32 of the original feature map is the lowest resolution available for the feature map. The basic block of ResNet-50 is used for forward propagation of the branch feature map, and its step is set to 1. At the same time, the feature maps between different branches always maintain information interaction. The convolutional operation with a step of 2 is still used from high resolution features to low resolution features, and bilinear interpolation for up-sampling is used for the conversion from low resolution to high resolution. After receiving information from different branches at each layer, the information is spliced to complete information fusion.

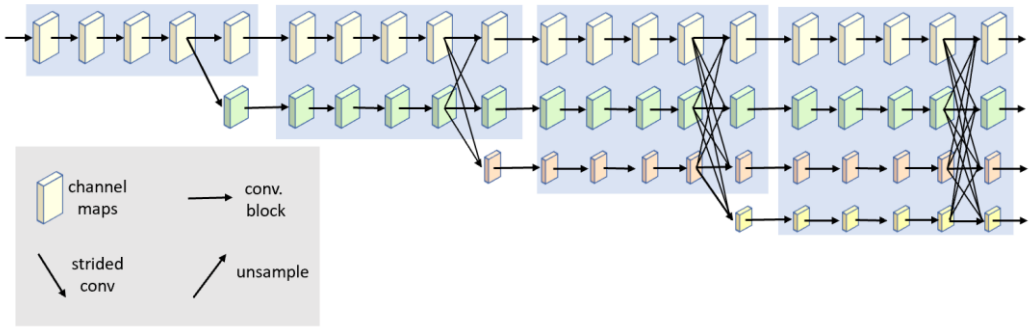


Fig. 3. The fast high resolution network.

There are a lot of feature convolutional operations and feature fusion in different branches of HRNet at different stages, so a large number of parameters and calculations will be generated. Too many residual modules bring a lot of calculations, and the most used operator in the residual module is 1×1 convolution. To address the issue of excessive computation, the residual module is enhanced by utilizing depthwise separable convolution. Each convolutional core of the depthwise separable convolution corresponds to only one channel, which can be used to replace 1×1 convolutions. Unlike the standard convolutional layer, the depthwise separable convolutional layer does not adhere to the size of $n \times n \times c$ for

the standard convolution operator; instead, it splits the convolutional kernel into c filters.

The depthwise separable convolution includes two steps: 1) Each channel of the h×w×c feature map undergoes a convolutional operation to produce c maps with dimensions of h×w×1. These maps are then stacked to form the h×w×c map, which serves as the input for the next convolutional layer. 2) m convolution blocks with the size of 1×1 are used to expand the depth of output, and finally the feature map with m channels is obtained. The batch normalization can not only make the network with higher learning performance, but also greatly accelerate the training process and achieve higher speed. Therefore, following each convolutional layer in the network, the batch normalization is incorporated to enhance to optimize of the training process.

3.3. The multi-scale feature aggregation module

As the backbone network, the HRNet maintains high resolution feature information. However, the convolution operation will lose some detail information during the process of downsampling and extracting deep high-level features. So, a novel multi-scale feature aggregation module is designed for building a dilated convolution spatial pyramid pooling module. The pyramid convolution (PyConv) operation can process the input data at multiple different filter scales. The PyConv contains a core pyramid. In order to learn features of different scales, each layer contains filters of different scales. Compared with the standard convolutional operation, the PyConv does not add additional parameters. The PyConv can capture information in different environments and scenes through multiple filters. In addition, the pyramid network enriches the scales for feature extraction, enhances the representation ability, and further improves the prediction accuracy. To compensate for the information loss resulting from the down-sampling strategy, the spatial pyramid pooling is used in the fast HRNet network. It uses the dilated convolution and the pyramid structure to aggregate the multi-scale features. The dilated convolution is to insert a null value among the parameters of the standard convolution. It expands the receptive field to keep the parameter quantity of the standard convolution unchanged, so that each convolutional core can obtain a wider range of information, avoiding spatial information loss induced by the pooling operation.

Finally, the feature maps of different resolutions are sent to the multi-scale feature aggregation module. As illustrated in Fig. 4, the size of the feature maps after the first branch are transformed to the same size of the first feature map. After that, the feature maps are sent to the dilated convolution spatial pyramid pooling (ASPP). The final results are subsequently obtained through the convolutional operator and softmax activation function. The dilated convolution pooling pyramid uses a variety

of scales of dilated convolution and global average pooling, and integrates different receptive fields to extract more effective information.

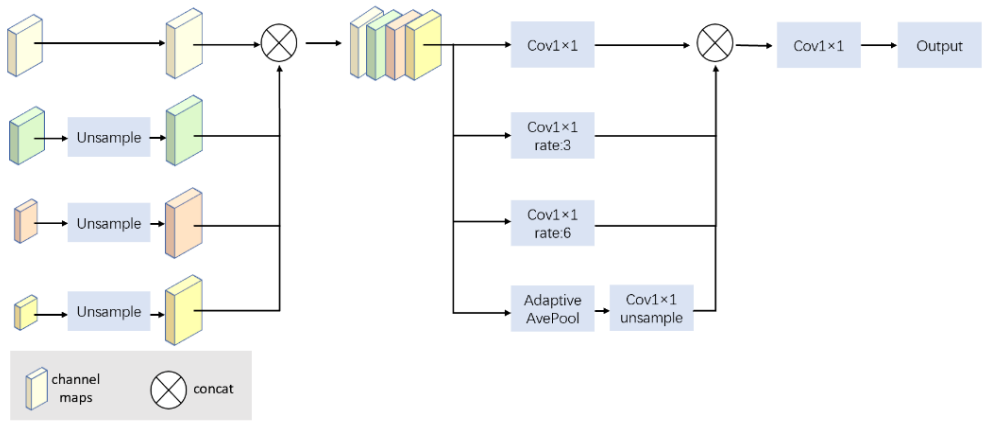


Fig. 4. The multi-scale feature aggregation module.

4. Experimental results

In this paper, our own coal and rock image segmentation dataset is built for evaluating the proposed segmentation method. The dataset has four categories: coal layer, roof layer, floating coal and background (the region between the coal layer and the roof layer, and the edge of image). The dataset contains 1763 images in total. Among them, 1243 images are used as the training samples, 161 images are used as the validation samples, and 359 images are used as the testing samples. Fig. 5 shows some examples of the coal and rock image segmentation dataset.

Fig. 5 Examples of the coal and rock image segmentation dataset.

Our experiments are carried out on NVIDIA GeForce RTX 2080Ti. The programming language is Python. For network training, the Adam optimizer is employed to adapt and constrain the learning rate. The batch size is 18. The training times is 300. The mean pixel accuracy (MPA), the mean intersection over Union

(MIoU) and the pixel accuracy (PA) are used for performance evaluation. In addition, the parameters of the network and FLOPs segmentation are utilized to evaluate the time efficiency.

To assess the effectiveness of the proposed coal and rock segmentation method, several other image semantic segmentation methods are performed for comparison [13,14]. It is can been seen from Table 1 that the parameter quantity of the original HRNet [1] is large, and it may not be suitable for practical applications. On the other hand, Lite-HRNet [11] significantly reduces computational requirements while maintaining similar segmentation performance to HRNet. In order to further enhance segmentation accuracy and speed, the FM-HRNet is designed. The number of repetitions of each phase in the basic network respectively are (1,1,4,3). First, the structural pruning is used to make the number of repetitions in each phase of the network become (1,1,1,1). At the same time, the number of repetitions of each block is reduced to 1. The above method is called the FM-HRNet-V0. Compared with HRNet, Lite-HRNet, Deeplabv3+, FCN-8s, U-Net and FAM-CRFSN[13], the FM-HRNet-V0 has higher segmentation accuracy and faster processing speed, but it still cannot be applied in the video surveillance system. So, the segmentation method FM-HRNet-V0 is further improved, and proposed FM-HRNet-V1. Subsequently, the method is further refined by downsampling input data and reducing channels in the last branch from 512 to 256 to decrease parameter count. According to Table 1, While FM-HRNet-V1 exhibits slightly lower segmentation performance than FM-HRNet-V0, it offers significant advantages in terms of segmentation speed. Therefore, the FM-HRNet-V1 is used as the basic network for subsequent work in this paper.

Table 1

Comparison of performance of four methods.

Model	MIoU	PA	MPA	FPS	GFLOPs
HRNet	86.74	93.48	82.59	6.7	174M
Lite-HRNet	86.94	93.59	92.16	10.4	21M
Deeplabv3+	87.15	94.23	91.65	7.8	135M
FCN-8s	84.32	92.85	87.58	11.5	83M
U-Net	86.87	93.17	89.36	12.6	52M
FAM-CRFSN	89.19	94.26	**94.10**	17.4	26M
FM-HRNet-V0	**89.28**	**94.90**	94.00	18.6	21M
FM-HRNet-V1	87.93	93.94	92.72	34.8	6.64M

Next, the segmentation performance after adding the depthwise separable convolution and the multi-scale feature aggregation module is investigated. Table 2 illustrates the performances of four different versions. The FM-HRNetFM-HRNet-V1 is an improved version after down-sampling the input feature maps based on the original HRNet. The FM-HRNet-V2 is an improved version after using the multi-scale feature aggregation module based on the FM-HRNet-V1. The FM-HRNet-V3 is an improved version after using the depthwise separable convolution

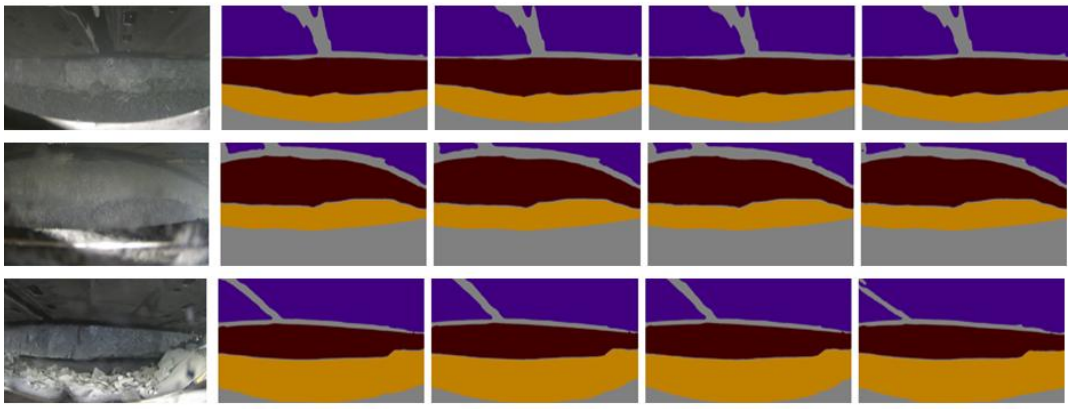
with the FM-HRNet-V1 method. The FM-HRNet-V4 is an improved version after using both the depthwise separable convolution and the multi-scale feature aggregation module with the FM-HRNet-V1. The experimental results of four versions are given in Table 2. The segmentation performance of FM-HRNet-V2 is better than that of FM-HRNet-V1, while the number of network parameters isn't changed. This suggests that the multi-scale feature aggregation module has little impact on reasoning speed but can enhance network segmentation accuracy. A comparison of experimental results between FM-HRNet-V1 and FM-HRNet-V3 reveals that depthwise separable convolution effectively improves reasoning speed. The best performance in both segmentation accuracy and reasoning speed is achieved by FM-HRNet-V4. Therefore, combining depthwise separable convolution with the multi-scale feature aggregation module effectively enhances coal and rock image segmentation performance.

To visually demonstrate the image segmentation capability of the network, four typical testing images are selected from the coal and rock image dataset. The original images and segmentation results for FM-HRNet-V1 to V4 are presented in Fig. 6. Additionally, Fig. 7 displays the coal and rock image segmentation results using FM-HRNet-V4, which clearly shows more distinct boundaries between different regions compared to other versions, as well as more accurate segmentation of image details.

Table 2

The different experimental results of four versions.

Model	MIoU	PA	MPA	Parameters	GFLOPs
FM-HRNet-V1	87.93	93.94	92.72	22.67M	6.64M
FM-HRNet-V2	89.23	94.88	93.99	22.67M	6.64M
FM-HRNet-V3	87.46	93.98	92.97	14.54M	5.38M
FM-HRNet-V4	89.26	94.88	93.99	14.54M	5.39M



(a) (b) (c) (c) (e)

Fig. 6. Segmentation results for four different versions. (a) Original images (b) FM-HRNet-V1. (c)FM-HRNet-V2. (d) FM-HRNet-V3. (e) FM-HRNet-V4.

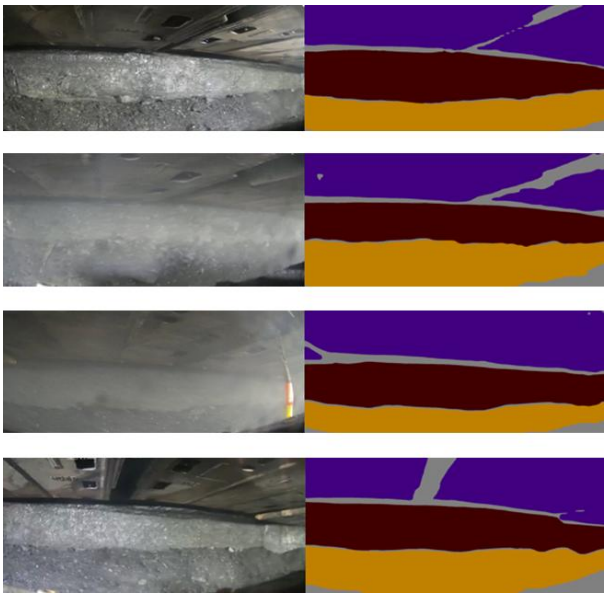


Fig. 7. Image segmentation results of FM-HRNet-V4.

5. Conclusions

This paper introduces a multi-scale HRNet for the rapid and accurate segmentation of coal and rock images. It achieves faster reasoning speed and has higher segmentation accuracy by enhancing the extraction of coal-rock layer features. Firstly, it effectively suppresses irrelevant feature information, while also mitigating gradient disappearance caused by excessive neural network layers and reducing network parameter count. The optimized pyramid convolution expands the model's receptive field, capturing diverse scene information through multiple filters to significantly enhance performance. Compared with HRNet and Lite HRNet, the proposed method improves segmentation accuracy and reduces parameters effectively. However, further enhancements are needed for more complex images. In the future, attention mechanisms should be explored to improve segmentation quality and computational efficiency.

Acknowledgments

This research was funded by the Key Science and Technology Innovation Project of China Coal Technology & Engineering Group Corp (No. 2022-3-KJHZ005).

REFERENCES

[1]. *K. Sun, B. Xiao, D. Liu and J. Wang*, "Deep high resolution Representation Learning for Human Pose Estimation", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5686-5696.

[2]. *Farabet C, Couprie C, Najman L, et al*, "Learning Hierarchical Features for Scene Labeling", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, 2013, pp. 1915-1929.

[3]. *Ghiasi G, Fowlkes C C*, "Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation", in European Conference on Computer Vision (ECCV), Amsterdam, Netherlands: Springer, 2016, pp. 519-534

[4]. *Chen L-C, Papandreou G, Kokkinos I, et al*, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, 2018, 834-848.

[5]. *Zhao H, Shi J, Qi X, et al*, "Pyramid Scene Parsing Network", in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii: IEEE, 2017, pp. 6230-6239.

[6]. *Fu J, Liu J, Tian H, et al*. "Dual Attention Network for Scene Segmentation", in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, California: IEEE, 2019, pp. 3146-3154.

[7]. *Chen L-C, Papandreou G, Kokkinos I, et al*, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs", in International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015, pp.1-14.

[8]. *Chen L-C, Papandreou G, Kokkinos I, et al*, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs", in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, vol. 40, no. 4, pp. 834-848.

[9]. *Chen L-C, Papandreou G, Schroff F, et al*, "Rethinking Atrous Convolution for Semantic Image Segmentation", arXiv preprint arXiv:1706.05587. 2017.

[10]. *Chen L-C, Zhu Y, Papandreou G, et al*, "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation", in European Conference on Computer Vision (ECCV), Munich, Germany: Springer, 2018, pp. 801-818.

[11]. *Yu C, Xiao B, Gao C et al*, "Lite-HRNet: A Lightweight high-resolution Network", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 10435-10445.

[12]. *Xu S, Jiang W, Liu Q et al*, "Coal-rock interface real-time recognition based on the improved YOLO detection and bilateral segmentation network", in Underground Space, 2024, in press, doi: https://doi.org/10.1016/j.undsp.2024.07.003

[13]. *Sun C, Li X, Chen J et al*, "Coal-Rock Image Recognition Method for Complex and Harsh Environment in Coal Mine Using Deep Learning Models", in IEEE Access, 2023, vol. 11, pp. 80794-80805.

[14]. *Sui Y, Zhang L, Sun Z et al*, "Research on coal and rock recognition in coal mining based on artificial neural network models", in Applied Sciences, 14(2), 864.