RailVID: A Dataset for Rail Environment Semantic

Hao Yuan
School of Rail Transportation
Soochow University
Suzhou, Jiangsu, China
20205246017@stu.suda.edu.cn

Zhenkun Mei
Suzhou Rail Transit Group Co.Ltd.
Suzhou, Jiangsu, China
meizk@163.com

Yihao Chen
School of Rail Transportation
Soochow University
Suzhou, Jiangsu, China
20204246013@stu.suda.edu.cn

Weilong Niu
School of Rail Transportation
Soochow University
Suzhou, Jiangsu, China
wlniu@stu.suda.edu.cn

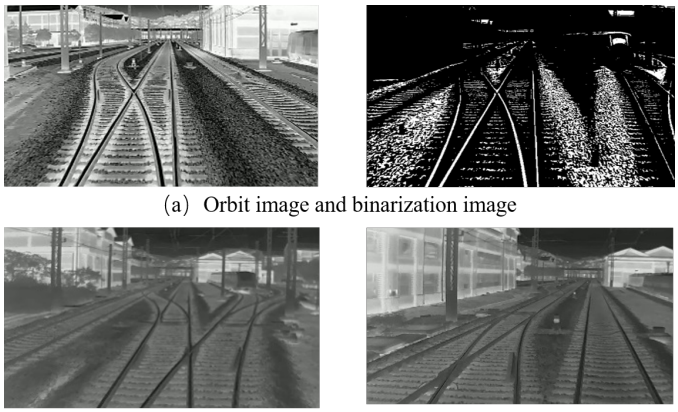
Cheng Wu
School of Rail Transportation
Soochow University
Suzhou, Jiangsu, China
cwu@stu.suda.edu.cn

***Abstract*—At present, rail transit is becoming the main means of urban and intercity fast passenger and freight transportation. Its safe operation is of great significance to protect the life of people, security of property and maintain social stability. Though the degree of intelligent traffic has been improved , there are still many safety risks in the current system by using manual hazard monitoring in the railway. Limited by the particularity and complexity of railway scenes, there are few studies on the perception and understanding of rail transit environment. In this paper, we propose a new rail transit dataset – RailVID. We use the AT615X Infrared thermography from InfiRay to collect data and record different railway scenarios, including carport, depot, and straight. We then propose an improved BiSeNet real-time semantic segmentation network for evaluation. Based on this dataset, we carry out environment perception, environment understanding, and safety decisions on the track area in front of the train, and we propose a solution for fully automatic train operation of rail transit.The dataset we provide compensates for the infrared data that is not in the existing dataset, and our data covers special weather and various conditions. Experiments show that our method achieves a higher Mean Pixel Accuracy in the collected dataset, and the processing speed also meets the real-time requirement.**

Index Terms*—Semantic segmentation; Rail transit; Environmental perception.

I. INTRODUCTION

At present, with the continuous advancement of modernization [1], rail transit has become the main trunk line of public transportation and the main artery of passenger and freight transportation. All trains, however, locate themselves on the rails through a safety control system to ensure the running interval [2], in order to ensure the running safety. In this context, there is a lack of means to perceive and understand the environment. Therefore, when manual observation is limited or signal equipment is faulty, it is difficult to judge the environment ahead, which may lead to serious accidents. In recent years, research on the environmental perception of urban rail transit has become a top priority. At the same time, the successful application of environmental perception in the intelligent assisted driving system of vehicles provides new ideas and solutions to the problems of environmental understanding and environmental perception of urban rail transit.



(a) Orbit image and binarization image

(b) Track lines are more complex than lane lines

Figure 1. Unique problems of track circuit.

In the current period of rapid development of urban rail transit, it has become the mainstream research direction to improve the intelligent level of operation with advanced scientific and technological methods. With the rapid development of deep learning and the progress of sensor technology, the application of deep learning in environment perception is being widely used in rail transit. Its main application fields are turnout fault diagnosis [3], passenger flow prediction [4], carriage security detection [5], train environment perception, etc. Deep learning has great potential in train environment perception. Through the deep learning algorithm processing of the data collected by the video sensor, the driver can improve the warning messages of road conditions ahead, which can effectively reduce the probability of accidents, ensure the safe operation of the train and improve the operation efficiency.

Image semantic segmentation refers to the specific pixel level when recognizing the image, that is, to divide each pixel of the image into a category, which is of great significance

 ISBN: 978-1-61208-941-6

in the environmental perception and understanding of rail transit [15]. However, in the case of poor imaging conditions such as night, rain, fog, and point light interference, semantic segmentation using visible images will lead to a decrease of visual perception. Infrared cameras use thermal radiation to generate thermal images, which are independent of the light source, are less affected by weather, and have longer detection distance. With an infrared camera, images are clearer and more stable at night when there is dark, rain, fog, or point light source pollution [16]. Using infrared thermography for real-time semantic segmentation of track environment can perceive the environment under the condition of poor visible light imaging, enhance the adaptability to special environment and special conditions, and improve the safety and intelligence of fully-automatic train operation. The unique problems of track circuit are shown in Figure 1.

The rest of the paper is structured as follows. In Section II, we present the current development of semantic segmentation and our contributions. In Section III, we introduce the BiseNet network which is suitable for infrared data, and propose an improved BiseNet to enhance the accuracy and meet the real-time demand. In Section IV, we introduce the equipment we used and the data collection scenarios, then we annotate and categorize the data. In Section V, we introduce the allocation and evaluation index of our model. Then, we compare the results of different methods on our dataset and show that our method is better. Finally, we conclude our work in Section VI.

II. Related work

Real-time semantic segmentation technology is mainly used in autonomous driving, which can help vehicles enhance their perception of their surroundings and understand their environment. In 2014, Long et al. [3] proposed a Fully Convolution neural Network (FCN), which replaced the full connection layer in the mainstream classification network with the convolution layer, and upsampled the feature map of the last volume base layer by restoring it to the same size as the input image, so that a prediction can be generated for each pixel. It preserves the spatial information in the original input image, which makes a breakthrough in semantic segmentation. Image semantic segmentation technology based on deep learning has begun to receive wide attention in the academic circle. SegNet [6] introduces the encoder-decoder concept for semantic segmentation. It uses the VGG (Visual Geometry Group) network as the encoder, and the decoder is symmetrically designed. The maximum pooled index at the corresponding encoder layer is called for upsampling. Chen et al. proposed the encoder-decoder with Atrous Separable for Semantic Image Segmentation to improve the Semantic Segmentation accuracy network. The precision of semantic segmentation in autonomous driving environment is further improved by an encoding and decoding structure. Fisher et al. [7] proposed the concept of Dilated Convolution for the loss of location information at the pooling layer, and ensured network accuracy with a larger receptive field. On this basis, Chen et al. proposed DeepLab series [8]–[10], using fully connected Conditional Random field (DenseCRF) and Atrous Spatial Pyramid Pooling (ASPP), respectively. The subsequent DeepLab v3+ [11] proposed a simple and effective decoder module based on the powerful encoder of DeepLab V3. While the DeepLab family improves network performance, it also increases computing costs. The Context Grided Network (CGNet) proposed by Wu et al. [12] uses lightweight framework and context joint feature extraction to improve the real-time performance of semantic segmentation. Lin et al. [17] proposed RefineNet network, which uses residual links to explicitly combine each sub-sampling layer with the following network layer to reduce memory usage and improve feature fusion between modules. Bilateral Semantic segmentation Net (BiSeNet) proposed by Yu et al. [13] adopts a bilateral structure of spatial path and context path, which gives consideration to semantic segmentation accuracy and real-time performance. At the same time, its simple network structure is conducive to later optimization.

Our contribution: Compared with automatic driving semantic segmentation technology of urban roads, the semantic segmentation in visual perception of rail transit has been developed later, and the environment of rail transit is more complex. Moreover, due to the particularity of rail transit, it is impossible to flexibly deploy platforms to achieve large-scale measurement and improvement in public sections. However, the perception and understanding of railway environment requires a large number of actual data to develop, test and verify the algorithm. Few rail transit datasets are available at present, and only RailSem19 is publicly available. Real-time semantic segmentation of railway environment using infrared thermography mainly involves the following four problems: 1) The rail transit environment is complex. 2) Infrared thermal images have more noise, low resolution, contrast and signal-to-noise ratio, and lack color features. 3) The semantic segmentation network model has a large amount of computation, which makes it difficult to meet the requirements of real-time. 4) An infrared dataset for the rail transit environment is lacking.

To solve these problems, we used the infrared thermography to collect the infrared data of Suzhou Rail Transit Line 1 in Jiangsu Province, China and proposed an infrared dataset for the rail transit – RailVID. By combining infrared thermal imaging and semantic segmentation based on deep learning, a real-time semantic segmentation method for rail transit is proposed. Keeping in mind the requirement of infrared image noise, low resolution, lack of color features, and real-time performance of the rail transit, we propose an improved BiSeNet, which is based on BiSeNet. The network is improved according to the characteristics of infrared images. The improved methods include: 1) The global maximum pooling layer is more effective in retaining the image texture information than the global average pooling layer. Therefore, the global maximum pooling layer is used to replace the global average pooling layer in the attention enhancement module and feature fusion module of the network to retain the texture details of infrared images. 2) Further fusion of the infrared image low-level features on the path of obtaining spatial information.

 ISBN: 978-1-61208-941-6

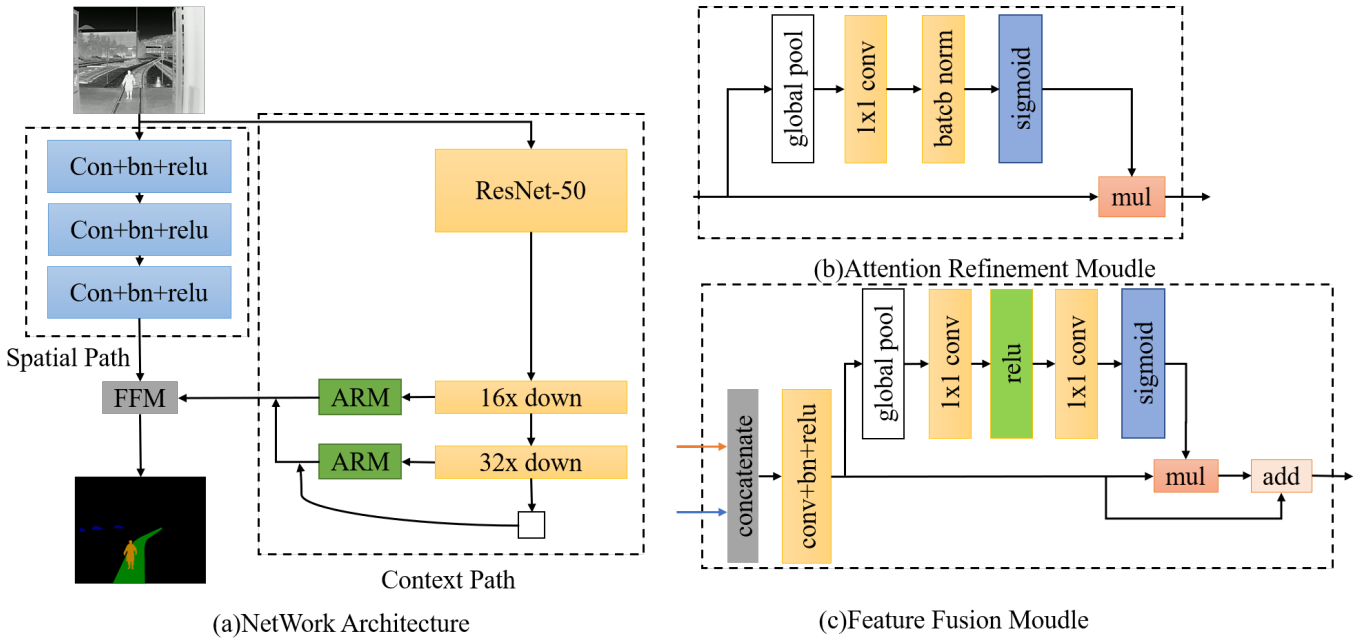


Figure 2. BiSeNet network structure.

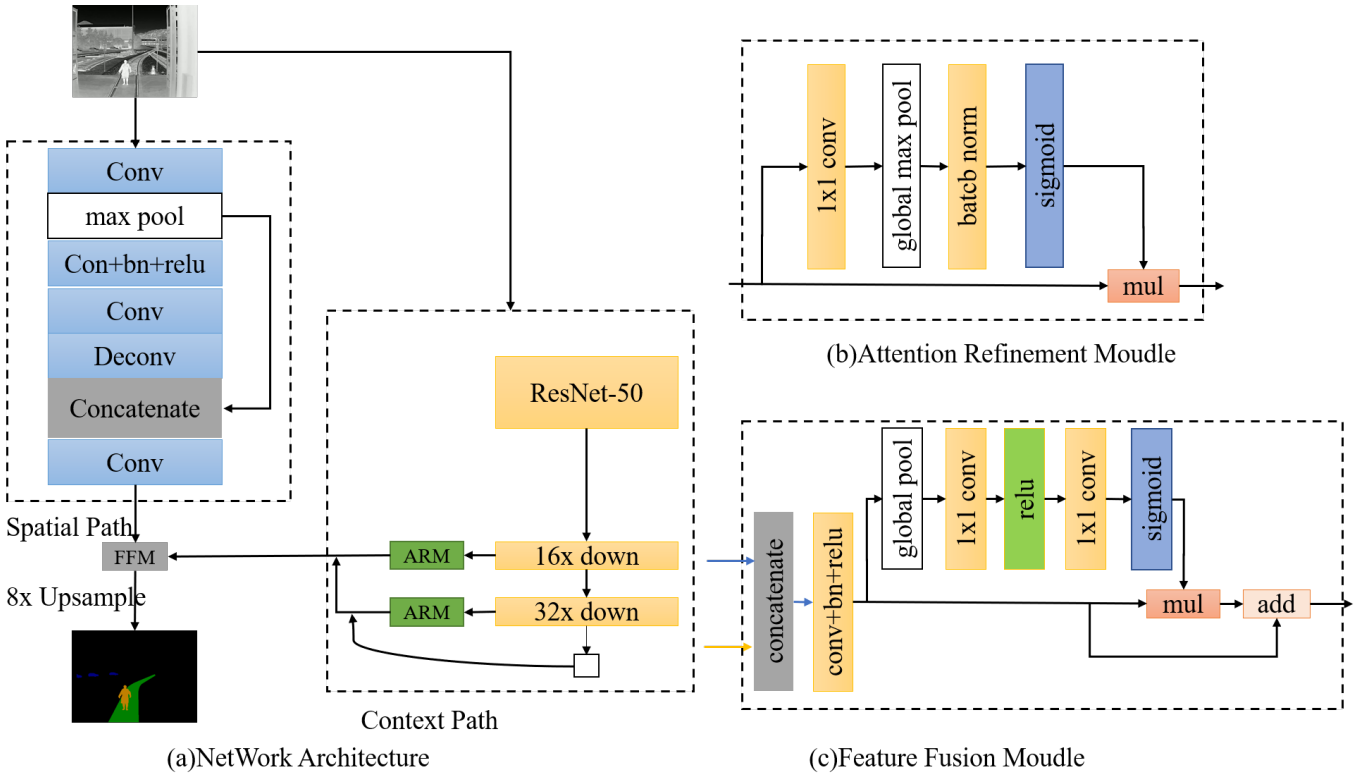
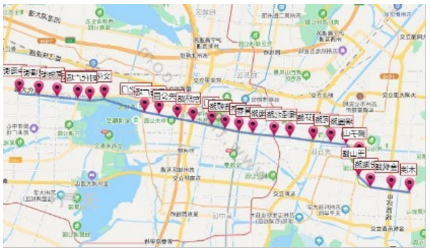


Figure 3. Improved BiSeNet network architecture.

III. Infrared Semantic Segmentation Network

A. *BiSeNet*

BiSeNet network structure is shown in Figure 2. In order to achieve fast real-time segmentation without sacrificing Spatial information, the BiSeNet structure is divided into two branches: Spatial Path (SP) and Context Path (CP). The spatial path uses three convolution layers with step 2, with batch normalization and activation function Relu to output 1/8 size feature maps of the original image, and uses feature maps of

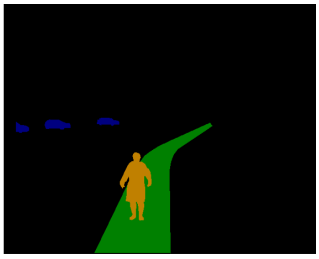


(a) Subway map

(b) Subway moving platform

Figure 4. RailVID data source and acquisition platform.

Original data



Label image

Figure 5. Samples of track area and surrounding obstacles.

a larger scale to retain rich spatial information. The context path uses a lightweight model Xception to extract high-level features and adds an attention optimization module at the end of the model. The upsampling operation is omitted, and the global average pooling is directly used to capture global context information and to calculate the attention vector to obtain a maximum perception field. Finally, we use the feature fusion module to fuse the low-level information in the spatial path with the high-level information output by the context path. The specific step is to join the two output features and then normalize the scale of the balance features. The output results are transformed into feature vectors through a global pooling operation, and then a weight vector is obtained by 1×1 convolution calculation, so that the features of the two can be re-weighted.

B. Improved BiSeNet

Based on the real-time bilateral-semantic segmentation network structure, we propose an improved bilateral-semantic segmentation network structure for rail transit infrared image semantic segmentation. The structure is shown in Figure 3.

According to the idea of infrared image processing in reference [14], a method of fusing the low-level features of infrared image by adding a pool layer, deconvolution layer and full connection layer is proposed. Thus, the network can better restore the spatial resolution of infrared images and provide a larger visual field. It can enhance the attention model in the context path and the feature fusion module integrating two-way features of real-time bilateral semantic segmentation network. According to the characteristics of fuzzy details and the low contrast of infrared image, the global maximum pool layer is used to replace the global average pool layer in each pooling layer of the module network architecture to retain the texture details of the infrared image.

Considering the characteristics of infrared images and the real-time requirements of the system, the context path adopts the ResNet-50 feature extraction network, combined with global pooling. Then, it merges the intermediate results of ResNet-50(16x downsampling and 32x downsampling) and the output of global pooling.

TABLE I. CATAGORY AND CORRESPONDING RGB VALUES.

Number	Category	R	G	B
1	**background**	0	0	0
2	**railway**	0	128	0
3	**car**	0	0	128
4	**people**	64	128	0

TABLE II. NUMBER OF IMAGES IN THE TRACK AREA FOR EACH SCENE.

Route scene	Carport	Depot	Tunnel
Straight	0	55	26
Curve	16	180	22
Fork	0	165	0

TABLE III. NUMBER OF SURROUNDING OBSTACLES IN EACH SCENE.

Object scene	Carport	Tunnel
Obstacles	442	165

IV. DATASET

A. Data collection

Figure 4 shows that we use the AT615X infrared thermal instrument from InfiRay to collect data, with the highest resolution up to 640×512, and 10 fps (frame per second) real-time images can be output at this resolution. Through the program we develop, the sensor data can be transmitted to the database in real-time, so as to realize data analysis and storage. We select the video frames at the proportion of 1 frame in every 10 frames to obtain all image data, and then screen the redundant images and blind areas according to the scene and train line conditions. Finally, we collect a total of 1071 images.

B. Data annotation

Due to the particularity of railway environment and the characteristics of noise and low definition of infrared imaging, we focus on the simple track area, people, and cars around the track area, as seen in Figure 5. Therefore, we mark the track area and classify people and cars using Labelme tool for manual annotation. When labeling segmentation categories, semantic segmentation tasks require that masks of different colors should be assigned to different categories to be segmented. Therefore, RGB values corresponding to Mask labels of each category with different colors are shown in Table I.

 ISBN: 978-1-61208-941-6

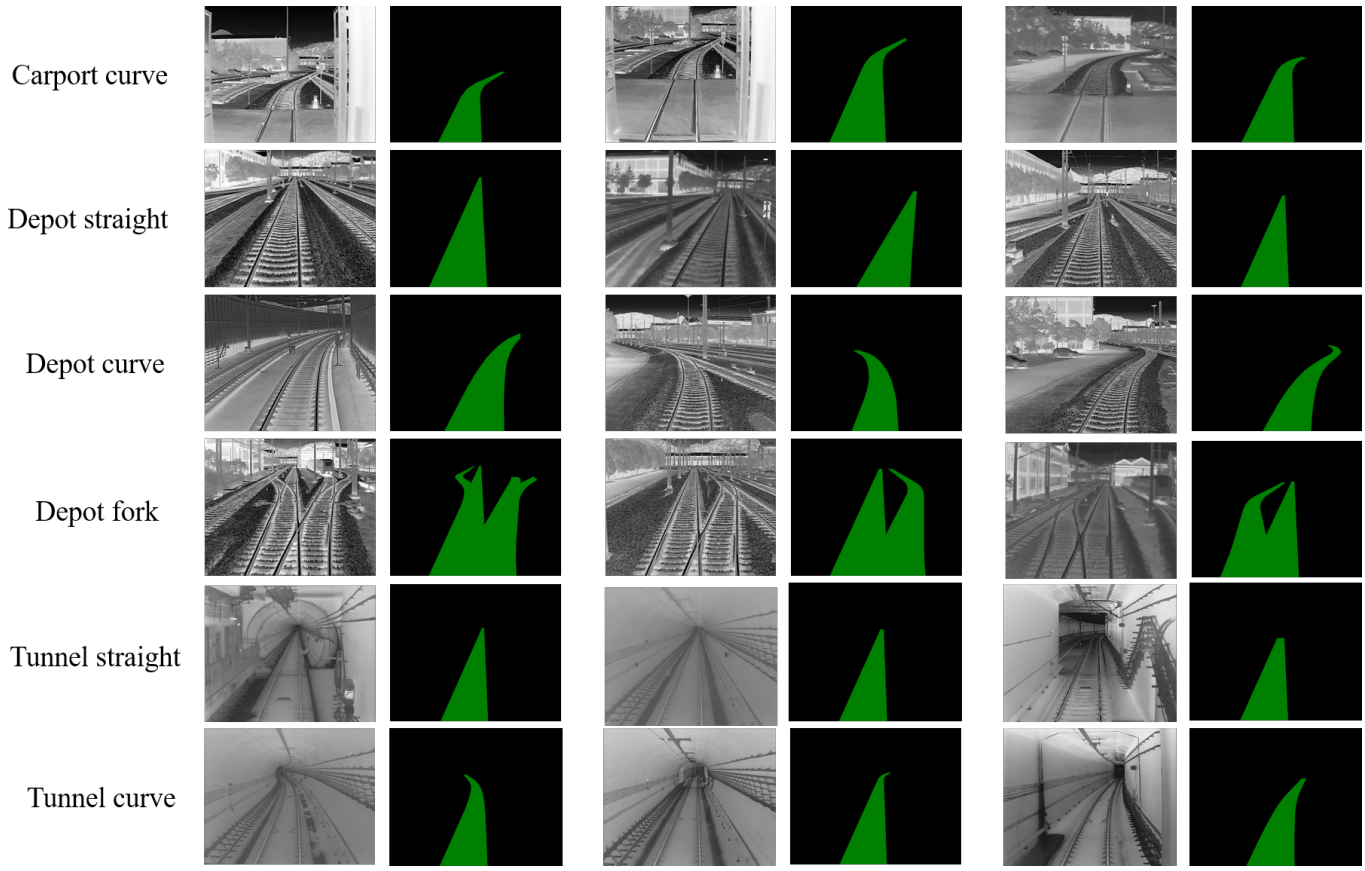


Figure 6. Track area original image and track area annotation example.

C. Data partitioning

According to the application scene, RailVID is divided into two parts: track area and obstacle, and a total of 1071 images are selected. Some examples of track area original images and track area annotation are shown in Figure 6. For the track area, we divide it into three parts: carport, depot and tunnel, and only mark the track area. The number of route scenes are shown in Table II.

There are many kinds of elements with high complexity in the railway environment. We also segmented other key elements to improve the perception ability of train perception and enrich the semantic information of railway scenes. Similarly, we divided the scene into two parts: carport and tunnel, and marked the track area, people and cars. The result are shown in Table III.

V. Experiment and Analysis

A. Model training

The specific software/hardware configuration of the improved RiSeNet we proposed is shown in Table IV. The ResNet-50 pre-training model was used to initialize some parameters. The RMSprop optimizer was used to optimize all parameters. The learning rate was set to 0.0001 and the attenuation rate was set to 0.995. The parameters of each layer of the deep network were randomly initialized. The training batch was set to 100, and the optimal semantic segmentation model was generated after training.

TABLE IV. HARDWARE AND SOFTWARE CONFIGURATION.

Configuration	Hardware / software
Platform	Win10
Environment	Tensorflow1.12.0+CUDA9.0
CPU	Core i7-8750H
GPU	Geforce GTX 950

B. Evaluation metrics

To verify the reliability of the method we proposed for infrared image segmentation of track area and surrounding areas, we assumed a total of $k+1$ classes (from L_0 to L_k, including an empty class background or a class). P_{ij} is the number of pixels that belong to class I but are divided into class J, P_{ji} is the number of pixels that belong to class J but are predicted to be class I, P_{ii} for real pixel number. The indicators used to evaluate real-time semantic segmentation are as follows:

1) Mean Intersection over Union ($mIoU$), the intersection of the predicted region and the actual region divided by the union of the predicted region and the actual region.

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ji}-P_{ii}} \quad (1)$$

 ISBN: 978-1-61208-941-6

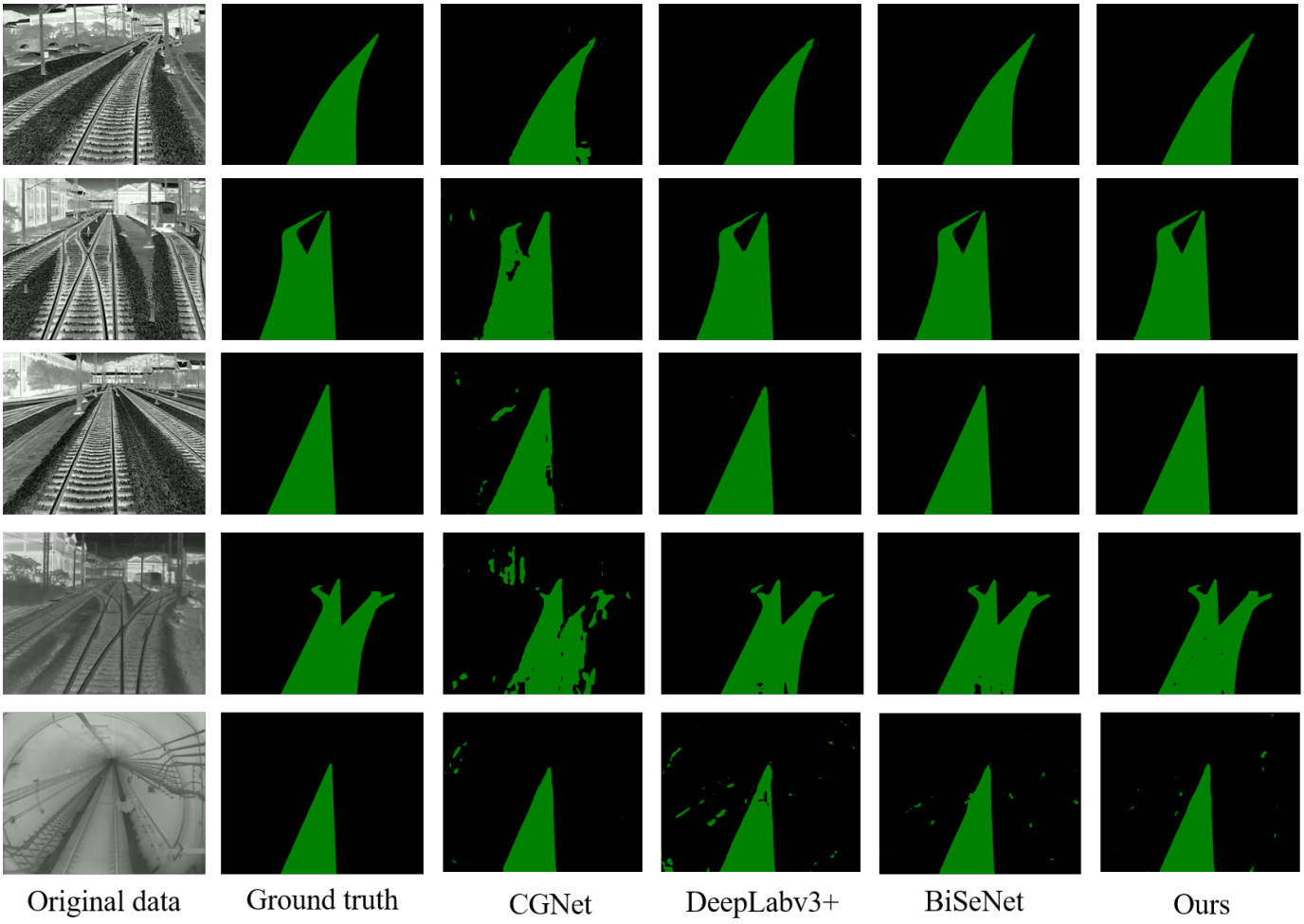


Figure 7. Comparison of effects of different semantic segmentation methods.

TABLE V. COMPARISON OF PERFORMANCE OF DIFFERENT METHODS.

Network	PA/%	mPA/%	mIoU/%	FPS/$f \cdot s^{-1}$
CGNet	60.33	51.17	59.24	53.0
DeepLabv3+	76.12	74.03	75.29	12.0
BiSeNet	78.01	76.54	77.57	45.2
ours	84.45	82.36	82.94	40.0

2) Pixel Accuracy (PA), the proportion of correctly labeled pixels to the total pixels.

$$PA = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} P_{ij}} \tag{2}$$

3) Mean Pixel Accuracy (mPA), the average Pixel Accuracy of all classes.

$$mPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij}} \tag{3}$$

4) Frames Per Second (FPS), which measures the real-time performance of the semantic segmentation algorithm.

$$FPS = \frac{N_f}{T} \tag{4}$$

In formula(4), N_f is the number of video frames, and T is the consumption time, in unit s.

C. Method comparison

To verify the advantages of the proposed method in infrared image segmentation of track area, we compared with the existing three representative semantic segmentation frameworks in visible environment (CGNet, DeeplabV3+, BiSeNet) over the same dataset with the same training parameters to test the models.

In the comprehensive test, infrared images of the track area taken by real shots in the test set were used for comparative testing. The comparison is made using pixel accuracy (PA), average pixel accuracy (mPA), Mean Pixel Accuracy (MIoU) and frame per second (FPS). The comparative test results

of the method we proposed compared with the other three methods are shown in Table V.

As can be seen from the comparative test index analysis in Table V, in terms of real-time performance, the improved BiSeNet is lower than BiSeNet due to the enhanced network depth. According to literature [4-5], if the FPS is above 25 frames /s, the algorithm can be identified as having real-time properties. Therefore, our method meets the real-time requirements of the semantic segmentation. The improved BiSeNet network, however, has obvious advantages over other three semantic segmentation methods in Pixel Accuracy (PA), Mean Pixel Accuracy (mPA) and Mean Intersection over Union ($mIoU$). For example, compared with BiSeNet network, our method improves the PA by 6.44 $\%$, the mPA by 5.82 $\%$ and the MIoU by 5.37 $\%$.

The improved BiSeNet network is optimized for retaining infrared image features and can retain more details of infrared image semantic segmentation in the track area and its surroundings. By comparing the infrared image segmentation results of the track area environment in the test set and the other three methods, it can be seen from Figure 7 that the segmentation effect of the improved BiSenet network is better and closer to the real annotation image.

VI. CONCLUSION AND FUTURE WORK

This paper described in detail about RailVID, a new infrared image dataset for rail transit that focuses on the intelligence and autonomy of rail transit train operation. We believe it will drive further development in the field of fully automated rail transit train operation. At the same time, we propose an improved BiSeNet real-time semantic segmentation network by fusing the low-level features of infrared image by adding a pool layer, a deconvolution layer and a full connection layer and replacing the global average pool layer with the global maximum pool layer for rail transit in the complex rail transit environment. On the collected dataset, the method achieves 82.94$\%$ of $mIoU$ and 40 f/s of FPS on the collected dataset, which satisfies the real-time semantic segmentation of rail transit. In the future, we will make more optimizations for the dataset and expand more data types, while a more efficient semantic segmentation network will be used to improve the characteristics of infrared images to achieve a better real-time semantic segmentation effect. This research is committed to providing the best method for fully automatic train operation in the rail transit scene and finally completing the construction of the feasibility platform.

ACKNOWLEDGEMENT

This work is supported by Scientific research project funding of Suzhou Rail Transit Group Co., Ltd. : SZZG06YJ6000017, thanks to corresponding author Cheng Wu, let me have the opportunity to participate in the meeting. Yuan and Mei contribute equally to the article.

REFERENCES

[1] B. Yum, "A Study on Railway Safety System Based on Accident Analysis," Journal of The Korean Institute of Plant Engineering, vol. 20, no. 1, pp. 101–106, 2015.

[2] J. Xiao, "Analysis of traffic accidents in Urban Rail Transportation," Information Week, vol. 1, no. 6, pp. 71–71, 2019.

[3] J. Long, E. Shelhamer, and F. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640–651, 2015.

[4] Z. Zhao, W. Chen, X. Wu, P. Chen, and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," LET Intelligent Transport Systems, vol. 11, no. 2, pp. 68–75, 2017.

[5] U. Safder, K. Nam, D. Kim, S. Heo, and C. Yoo, "A real time QSAR-driven toxicity evaluation and monitoring of iron containing fine particulate matters in indoor subway stations," Ecotoxicology and Environmental Safety, vol. 169(MARa), pp.361–369, 2018.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, 2019.

[7] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," Computer Science arXiv:1511.07122v1.

[8] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," Computer EnCE no. 4, pp. 357–361, 2014.

[9] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell vol. 40, no. 4, pp. 834–848, 2018.

[10] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," arXiv preprint arXiv:1706.05587, 2019.

[11] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," In: Proceedings of the 15th European Conference on Computer Vision, pp. 833–851. Springer International Publishing, Switzerland, 2018.

[12] T. Wu and S. Tang, "CGNet: A Light-weight context guided network for semantic segmentation," IEEE Conference on Computer Vision and Pattern Recognition. Intell., vol. 30, pp. 1169–1179, 2019.

[13] R. Barth, J. Ijsselmuiden, J. Hemming, E. Henten, and J. Van, "Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation," Computersand Electronics in Agriculture, Intell vol. 161, pp. 291–304, 2019.

[14] Z. He, Y. P. Cao, Y. F. Dong, J. X. Yang, and Y. L. Cao, "Single-image-based nonuniformity correction of uncooled long-wave infrared detectors: A deep-learning approach," Applied Optics. Intell., vol. 57, no. 18, pp. 155–164, 2018.

[15] Y. H. Chen, N. Zhu, Q. Wu, C. Wu, W. L. Niu, and Y. M. Wang, "MRSI: a multimodal proximity remote sensing data set for environment perception in rail transit," Int J Intell Syst, pp. 1–27, 2021.

[16] Z. Chen, C. Wu, L. J. Zhang, and Y. M. Wang, "Near Real-time Situation Awareness and Anomaly Detection for Complex Railway Environment," IEEE Conference on Cognitive and Computational Aspects of Situation Management, pp. 1–8, 2021.

[17] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5168–5177, 2017.

 ISBN: 978-1-61208-941-6