Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on semantic segmentation methods for RGB-D urban scenes in the context of artificial intelligence

Xiangling Ma[1], Xiangyang Ma[2,✉], Minghui Qiu[1]

[1] *School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou, Guangdong, 511363, China*

[2] *Human Resources Office, Shandong Jianzhu University, Jinan, Shandong, 250101, China*

ABSTRACT

To solve the problem of identifying intrinsic relationships between objects and mirror segmentation in semantic segmentation of urban scenes using current multi-modal data, this study innovatively integrates color images, depth information, and thermal images to propose a network model that integrates modal memory sharing and form complementarity, and a hierarchical assisted fusion network model. Compared with existing advanced urban scene semantic segmentation methods, the proposed method performed excellently in terms of performance, with an average pixel accuracy and mean intersection over union of over 80% for different objects. In addition, the research method achieved clearer and more complete segmentation results by strengthening contextual associations, and edge processing is also smoother. Even in object segmentation with similarities in distance, shape, and brightness such as "vegetation" and "sidewalk", the research method still maintained high accuracy. The research method can effectively handle the complexity of urban scenes, providing a new solution for semantic segmentation of multi-modal data in urban scenes.

Keywords: RGB-D, knowledge distillation, modal adaptation, urban scenes, semantic segmentation

1. Introduction

The purpose of semantic segmentation task is to recognize the semantic categories of pixels in an image based on the content of the input image. As the foundation of intelligent scene understanding, semantic segmentation is of great research significance and is widely used in fields such as autonomous driving and robot perception [3, 16, 5, 10]. The earlier semantic segmentation of indoor scenes is

✉ Corresponding author.
E-mail address: 18615238728@163.com (X. Ma).

Received 11 September 2024; Accepted 02 February 2025; Published Online 16 March 2025.

DOI: 10.61091/jcmcc124-13

based on RGB images to calculate the semantic segmentation results, due to the influence of many object categories in the scene, serious mutual occlusion phenomenon between objects [12, 13, 15], light imbalance and other problems, it is difficult to accurately segment the complex scene only by unimodal RGB information, so the effect of semantic segmentation of indoor scenes based on RGB images needs to be further improved [18, 11, 17].

In recent years, with the emergence of depth sensors, RGB images containing texture information and Depth images containing depth information can be obtained synchronously, which promotes the research of semantic segmentation of indoor scenes based on RGB-D images, but the current RGB-D semantic segmentation methods still suffer from the problem of insufficient fusion of RGB and Depth information [4, 14, 6, 1]. RGB images are able to describe the appearance information such as color and texture of an object, and depth images can describe the spatial geometric information of an object. Compared with RGB images, Depth images provide light-independent geometric information, which helps to reduce the uncertainty of semantic segmentation caused by differences in lighting conditions [8, 2, 9, 7].

In summary, the current semantic segmentation methods mainly include deep learning methods, multi-resolution learning methods, and multimodal data methods, each of which has its own advantages. However, deep learning methods typically require mass annotated data for training, which can be expensive and time-consuming in practical applications. Although multi-resolution learning methods can improve detail recognition ability, they may increase computational complexity and affect real-time performance. Multi-modal data methods also face challenges in data fusion and handling inconsistencies. In view of this, taking urban street scenes as the research object, this study innovatively proposes a network model that combines modal memory sharing and form complementarity, as well as a hierarchical assisted fusion network model. These models utilize RGB-D and Thermal image data within a supervised learning framework to explore better solutions for semantic segmentation problems. This multi-modal data fusion and innovative network architecture design aims to further improve the performance of semantic segmentation while reducing dependence on a large amount of annotated data to meet real-time and robustness requirements.

2. Methods and materials

2.1. Construction of complementary network and modal memory sharing model for urban scene morphology based on RGB

The fusion algorithm based on multi-modal and multi-level features has shown great potential in computer vision. In order to effectively integrate multi-modal information, various fusion strategies have been proposed and have achieved certain results. However, these methods still face some challenges. The current main challenge lies in how to make the model not only limited to identifying and analyzing the features of individual samples, but also able to recognize the inherent connections between samples, and how to more effectively integrate feature information of different scales to optimize the performance and generalization ability. In order to address the challenges in urban scene analysis, a RGB-based urban scene form complementary network and modal memory sharing model, namely RGB-MMS-MCN model, is proposed. This model aims to enhance the accuracy and robustness of urban scene recognition by integrating multiple sources of data. The overall framework structure of the RGB-MMS-MCN model is shown in Figure 1.

As shown in Figure 1, the RGB-MMS-MCN model adopts an encoder-decoder structure to ef-

fectively process and analyze RGB and Thermal data in urban scenes. Among them, the encoder part is based on a backbone network (Segmentation Transformer, SegFormer), which extracts modal features by passing two types of data graphs into the RGB branch and Thermal branch, respectively. These features are then fed into the MMS to learn and integrate the bimodal information of RGB and Thermal maps from the entire dataset. The output of the MMS module integrates two modal features, and the output size of the first three MMS modules is half of the input. Therefore, in order to reduce the computational burden, the number of output channels for all four MMS modules is uniformly adjusted to 64. In the decoding stage, the model constructs an integrated decoding unit, which includes a Skeletal Pose Machine (SPM), a Convolutional Pose Machine (CPM), and an MCN. These modules work together to achieve precise identification and analysis of complex structures in urban scenes. In the MMS module, this study extracts combinations and commonalities of features within a range through global average pooling, while utilizing parallel branches to learn shared features. Finally, the features of the middle branch are combined with the fused features of the upper layer to achieve information enhancement. The contour positioning process of CPM module is shown in formula (1).

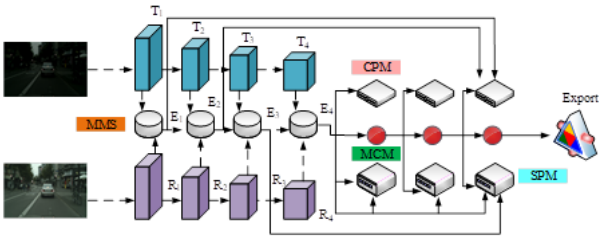


Fig. 1. The overall framework of the RGB-MMS-MCN model

$$\begin{cases} E_{mul} = CBL\left(E_1\right) \times CBL\left(Up_2\left(E_2\right)\right), \\ E_{cat} = Cat\left(CBL\left(E_1\right), CBL\left(Up_4\left(E_4\right)\right)\right), \\ C_1 = CBL\left(MHDC\left(CBL\left(E_{mul}\right) + CBL\left(E_{cat}\right)\right)\right). \end{cases} \tag{1}$$

In formula (1), E_{mul} and E_{cat} represent combined features. E_1, E_2, and E_4 represent input feature maps. C_1 represents the final output feature map, which is the concatenation result of E_{mul} and E_{cat} after multi-head expansion convolution processing, and then processed by convolution blocks. $MHDC$ represents multi-head expansion convolution. Up_i represents i-fold upsampling operation. The entire calculation process is a feature fusion and extraction process, which enhances the model's understanding of input data through different combinations of operations. The recognition and localization process of the SPM module target skeleton is shown in formula (2).

$$\begin{cases} E_{add} = Up_8\left(E_3\right) + Up_8\left(E_4\right), \\ E_{cat} = Cat\left(Up_8\left(E_3\right), Up_8\left(E_4\right)\right), \\ E_{enhance} = CCMP\left(E_{cat} \times Conv\left(E_{add}\right)\right). \end{cases} \tag{2}$$

In formula (2), E_{add} represents the feature map obtained by concatenating E_3 and E_4 after upsampling. $E_{enhance}$ represents the feature map obtained by element wise multiplication of E_{cat} after cross-channel max pooling and E_{add} after convolution. The structural flow of MCM module is shown in Figure 2.

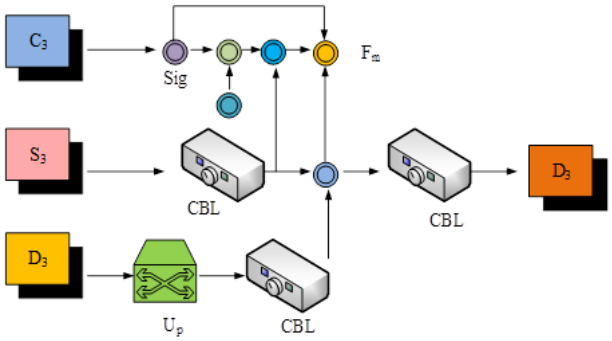


Fig. 2. The overall framework of the RGB-MMS-MCN model

As shown in Figure 2, the MCM module can extract and enhance additional features of contour information through a series of complex operations. Firstly, applying activation functions to introduce nonlinearity enables the model to learn more complex feature representations. Subsequently, by performing element by element subtraction, the key differences in the contour information are highlighted, thereby enhancing the discriminative ability of the features. Secondly, residual connections are used to further enhance contour information. Finally, the previously extracted complementary features, input skeleton features, and newly generated complementary features are integrated and convolved to produce the final complementary features. In this way, MCM modules can not only retain useful information from the original features, but also introduce new complementary information. The specific process of MCM module form combination is shown in formula (3).

$$\begin{cases} F_m = (1 - Sig\,(C_3)) \times CBL\,(S_3) + C_3, \\ D_3 = CBL\,(Cat\,(F_m, CBL\,(S_3)\,, CBL\,(Up_2\,(D_2)))). \end{cases} \tag{3}$$

In formula (3), C_3 and S_3 represent two different input feature maps. Sig represents the Sigmoid activation function. F_m is the feature map obtained by multiplying C_3 with the convolution block output of S_3 after passing through the Sigmoid activation function, and then adding C_3 element by element. The entire calculation process is a process of feature fusion and complementarity, which enhances and fuses features through operations such as Sigmoid activation function, element wise multiplication, element wise addition, upsampling, concatenation, and convolution blocks, ultimately obtaining complementary feature D_3.

2.2. Construction of a hierarchical assisted fusion network model for urban scenes based on RGB-D

Urban scene mirror segmentation refers to the process of segmenting various elements in a city, such as buildings, roads, vegetation, etc., and mirroring them to achieve specific visual effects. This technology can be applied in fields such as urban planning, architectural design, virtual reality, etc., helping designers and developers better understand and showcase the urban environment. Although depth information is crucial for accurate image segmentation, how to use RGB-MMS-MCN model to accurately distinguish between images and actual objects remains a challenge to be solved. To solve this problem, a novel RGB-D Cross-Modal and Progressive Feature Fusion Network (RGB-D-CM-PFN) is proposed based on the RGB-MMS-MCN model, as displayed in Figure 3.

As shown in Figure 3, the RGB-D-CM-PFFN model mainly has two stages, namely the encoding stage and the decoding stage. In the former, ConvNext is used as the backbone network and divided into a five layer structure. In the first four layers, the study utilizes the Cross-Modal Fusion with

Information Integration (CFI) module to fuse cross-modal information. At the same time, in the last layer of the encoder, the study also introduces a Multi-scale Depth Attention (MDA) module to enhance context awareness. In the decoding stage, the study adopts an inverted pyramid structure. Among them, the first layer is composed of the Neural Architecture Evaluation (NAE) module, which is responsible for describing the mirror target from different perspectives. The Cross-Channel Fusion (CCF) module in the second layer is responsible for cross-information processing. This design not only optimizes the feature fusion process, but also enhances the semantic segmentation ability of the model for complex urban scenes. The cross-modal information fusion process of CFI module is shown in formula (4):

$$\begin{cases} R_{ai} = (CA(R_i \times C_i) + Sig(Conv_{7.7}(SA(R_i)))) \times C_i, \\ D_{ai} = (CA(D_i) \times C_i + Sig(Conv_{7.7}(SA(D_i)))) \times C_i. \end{cases} \tag{4}$$

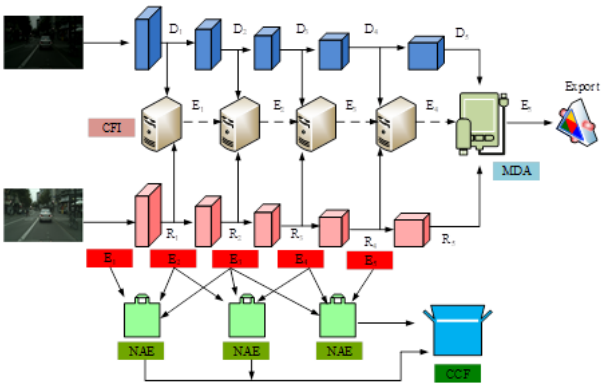


Fig. 3. Overall framework structure of the RGB-D-CM-PFFN model

In formula (4), CA and SA represent spatial attention and channel attention mechanisms, respectively, used to extract important information and global dependencies from feature maps. R_{ai} and D_{ai} represent two types of feature maps processed by spatial attention and channel attention mechanisms. $Conv_{7.7}$ represents 7×7 convolution operation. R_i, C_i, and D_i represent different features of the i-th channel. The semantic enhancement process of MDA module is shown in Figure 4.

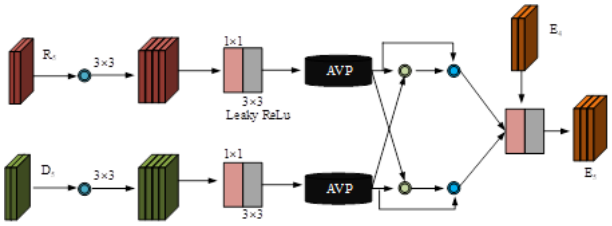


Fig. 4. Semantic enhancement process for MDA module

As shown in Figure 4, in order to enhance the model's understanding of environmental context, LDASPP and convolutions with different dilation rates (3, 6, 12) are used. To reduce the computational burden, the study also reduces the channel to 128 through convolution operations. Next, CBR convolutional blocks are used to capture detailed information. The two branches integrate information through cross-fusion operations. Finally, the number of channels is adjusted through convolution to match the original input, completing the refinement of the features. The NAE module describes the structural flow of mirror targets from different perspectives, as shown in Figure 5.

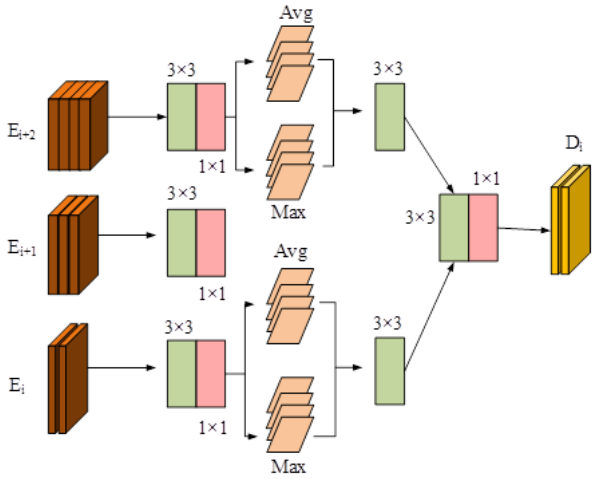


Fig. 5. The NAE module describes the structural flow of the mirrored target from different perspectives

As shown in Figure 5, the study first divides the input features into three levels, namely low, medium, and high. This hierarchical processing helps the model capture the features of data at different levels, thereby more effectively extracting and representing information. In addition, to reduce computational complexity, the study also uses a 1×1 convolution kernel to reduce the channel to 28 without increasing the number of parameters. The CBR module is used for precise feature extraction. In order to further enhance the expressive power of features, the study also adopts maximum pooling and average pooling techniques to extract two different spatial features. Maximum pooling can capture salient features in an image, while average pooling can capture the global information of the image. After fusing the features extracted by these two pooling techniques, the model can obtain richer spatial feature representations, enhancing the expressive power of the features. Finally, the enhanced spatial features are extracted through the CBR module and 1×1 convolution to optimize the feature representation. This optimized feature representation helps improve the model's understanding of image content, especially when dealing with complex visual tasks, providing more accurate predictions and classifications. In addition, during the training phase of the RGB-D-CM-PFNN model, three different loss functions are used to effectively supervise the multi-scale mirror output.

3. Results

3.1. Performance testing of urban scene hierarchical assisted fusion network model based on RGB-D

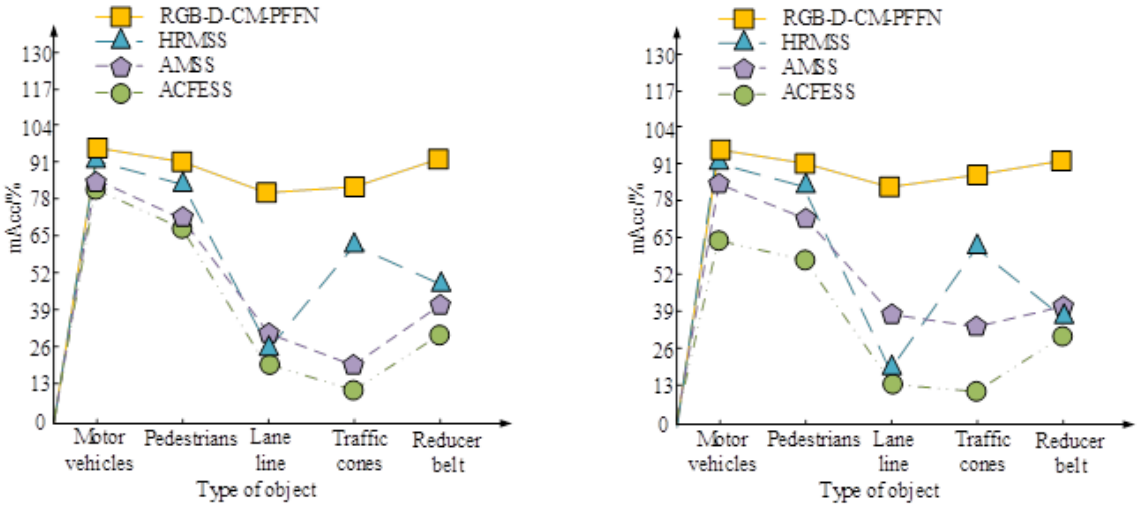
To verify the performance, a suitable experimental environment is established. The study uses Ubuntu 16.04 as the operating system, equipped with an Intel Core i7 CPU, NVIDIA GeForce GPU, 64GB of memory, and implements using the Pytorch framework. The urban landscape dataset Cityscapes and PST900 are used as data sources. Among them, Cityscapes is a large-scale dataset focused on understanding urban street scenes, containing images of 50 cities in different environments. The PST900 dataset is a dataset focused on multi-spectral RGB semantic segmentation, providing 894 pairs of synchronized and calibrated RGB and thermal imaging images. The study divides these two datasets into training and testing sets in an 8:2. Table 1 displays the specific parameter settings.

According to the parameter settings in Table 1, the study introduces popular urban scene se-

mantic segmentation models, namely Attention Mechanism-based Semantic Segmentation (AMSS), High-Resolution Maintenance Semantic Segmentation (HRMSS), and Atrous Convolutional Feature Extraction Semantic Segmentation (ACFESS). Firstly, a comparative experiment is conducted using Mean Pixel Accuracy (mAcc) as the testing metric, and the test results are shown in Figure 6.

Table 1. Experimental parameter setting

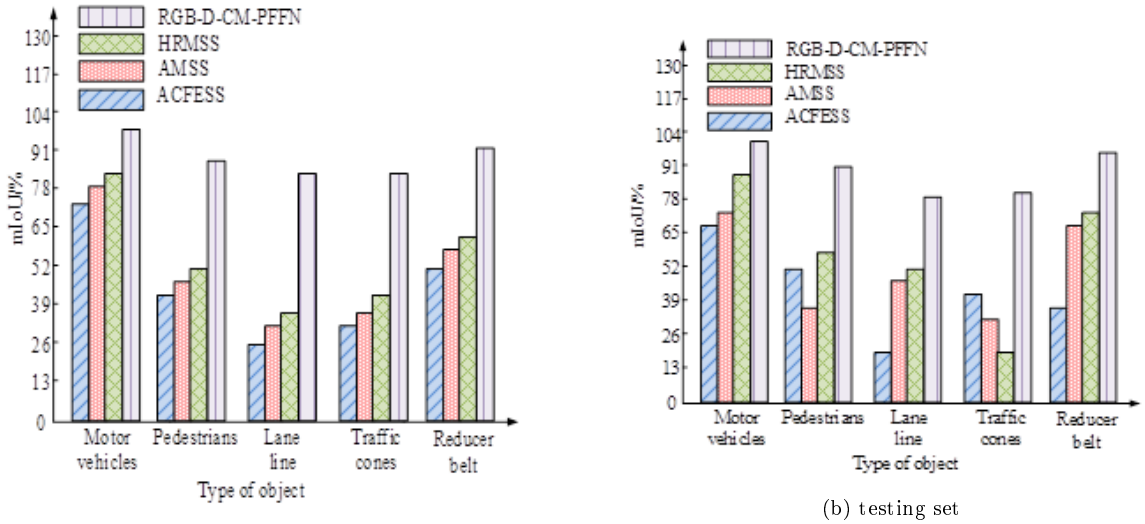
Serial number	Parameters	Settings
1	Backbone network	SegFormer-B3
2	Number of training sessions	200
3	Optimizer	Ranger
4	Batch size	2
5	Initial learning rate	1×10^{-4}
6	Weight decay	5×10^{-4}
7	Calculation of losses	Cross-entropy loss



(a) training set (b) testing set

Fig. 6. Comparison curve of average pixel accuracy for different models

Figures 6a and 6b show the mAcc test results on the training and testing sets, respectively. As shown in Figure 6, the RGB-MMS-MCN model achieved the best performance in segmenting all types of objects. This model performs outstandingly in integrating RGB information for feature extraction and fusion, effectively handling the complexity of urban scenes. The RGB-MMS-MCN model achieves segmentation accuracy of over 80% for all types of objects. Compared with other models, the performance improved by about 10% to 40%. The high performance of the model is attributed to several key factors. Firstly, the RGB-MMS-MCN model adopts an advanced network structure, which is highly suitable for processing high-resolution image data. Secondly, the model utilizes multi-scale feature fusion technology, which enables the model to capture both large-scale contextual information and small local details simultaneously. In addition, to determine the other performance indicators of the RGB-MMS-MCN, a comparative test is conducted on the four models using mIoU as the testing metric, as displayed in Figure 7.



(a) traning set

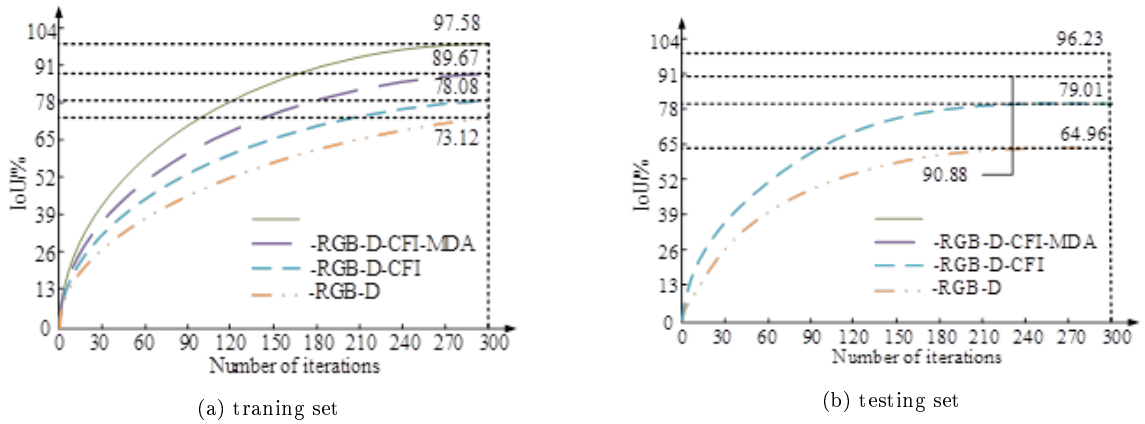
(b) testing set

Fig. 7. mIoU test results for different models

Figures 7a and 7b show the mIoU test results on the training and testing sets, respectively. As shown in Figure 7, the RGB-MMS-MCN model achieved mIoU of 97.03%, 88.62%, 82.16%, 82.22%, and 92.33% for motor vehicle, pedestrian, lane line, traffic cone, and reducer belt, respectively, which were significantly better than other models. This once again proves the efficient ability of integrating RGB and depth information for feature extraction and fusion. The combined effect of multi-modal information fusion, feature extraction network, optimized training strategy, and multi-scale feature fusion has enabled the RGB-MMS-MCN model to perform well in semantic segmentation tasks in urban scenes, especially in terms of segmentation details and edge accuracy.

3.2. Performance testing of urban scene hierarchical assisted fusion network model based on RGB-D

To verify the impact of each module in the RGB-D-CM-PFNN model on the overall performance, the study first conducts ablation tests using the Total Intersection over Union (IoU) as an indicator. The test results are shown in Figure 8.



(a) traning set

(b) testing set

Fig. 8. RGB-D-CM-PFFN model ablation test results

Figures 8a and 8b display the ablation test results in the training and testing sets, respectively. As the iteration increased, the IoU of each module in the RGB-D-CM-PFNN model gradually increased and then tended to balance. Both in the training and testing sets, the RGB-D module performed the worst, with a maximum IoU of only 69.28%. After improving through cross-modal fusion information integration, multi-scale deep attention, and supervision mechanisms, the IoU of the RGB-D module has increased by about 10% to 20%. The RGB-D-CM-PFFN model proposed in the study had the best comprehensive performance, with an optimal performance of 95.76% in semantic segmentation of urban scenes. In summary, the model significantly improves the IoU performance of semantic segmentation by integrating different modules. Finally, to demonstrate that the RGB-D-CM-PFNN model has better application capabilities in urban scene mirror segmentation compared with traditional semantic segmentation networks, visualization experiments are conducted on scene segmentation prediction results and segmentation edges. The test results are shown in Figure 9.

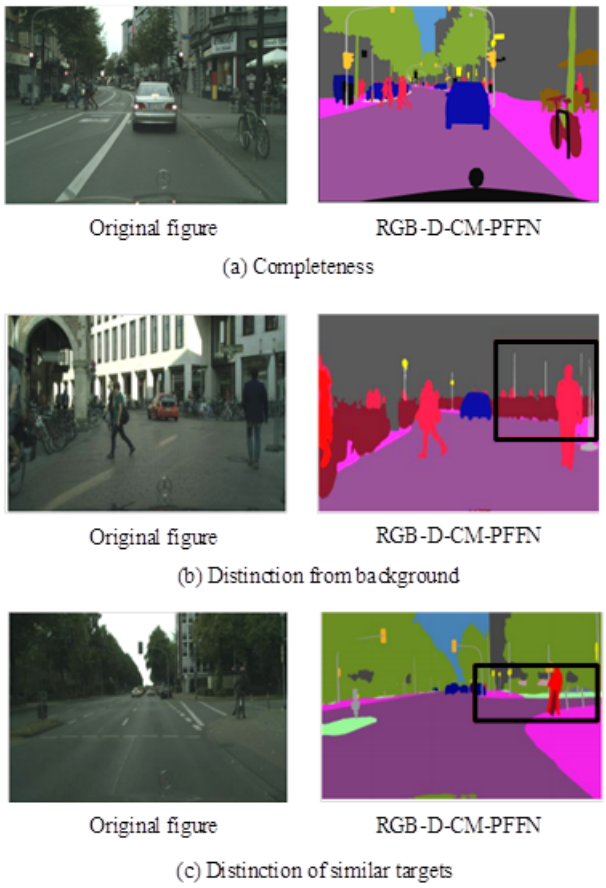


Fig. 9. RGB-D-CM-PFFN model visualization results

Figure 9 shows the performance of the RGB-D-CM-PFNN model in semantic segmentation under different scenarios. Figures 9(a), 9(b), and 9(c) present the experimental results of completeness, background discrimination, and similar object discrimination, respectively. As shown in Figure 9 (a), the RGB-D-CM-PFFN model achieved clearer and more complete segmentation results by enhancing context correlation, and the edge cutting was also smoother. As shown in Figure 9(b), the RGB-D-CM-PFFN model effectively distinguished target objects with high confusion with the background, such as road markers and utility poles. In Figure 9(c), although the "vegetation" and "sidewalk" were similar in distance, shape, and brightness, the RGB-D-CM-PFNN model still performed segmentation accurately. These results collectively demonstrate the effectiveness and superiority of the RGB-D-

CM-PFNN model in handling complex urban scenes.

4. Conclusion

With the rapid development of devices such as car cameras and surveillance cameras, a large number of urban street scene images have been generated. Analyzing the semantic information in these images is crucial for promoting the application of smart cities such as autonomous driving and intelligent services. However, traditional semantic segmentation methods overly rely on the selection of artificial features. Faced with the complexity of urban streets, it is often difficult to achieve the required high accuracy, which cannot meet the current high standards for image processing quality. In view of this, the study proposed a city scene form complementarity and modal memory sharing network model based on RGB and a city scene hierarchical assisted fusion network model on the basis of RGB-D image data. The experimental results showed that the mIoU of the research model for motor vehicle, pedestrian, lane line, traffic cone, and reducer belt in urban scenes reached 97.03%, 88.62%, 82.16%, 82.22%, and 92.33%, respectively. The segmentation mAcc for the above types of objects also reached over 80%. In addition, the research model effectively distinguished target objects with high confusion with the background, such as road markers and power poles. By enhancing contextual correlation, clearer and more complete segmentation results have been achieved, and edge cutting is also smoother. From this, the model has obvious advantages in processing urban scene images, especially in terms of segmentation details and edge accuracy. However, further exploration is necessary to improve the inference speed while enhancing its performance. In the future, lightweight network design can be adopted to improve the flexibility of the model.

Acknowledgements

The 14th-Five Year Plan Projects of Guangdong Association of Higher Education (22GYB159).

References

[1] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li. Shapeconv: shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7088–7097, 2021.

[2] F. Fooladgar and S. Kasaei. A survey on indoor rgb-d semantic segmentation: from hand-crafted features to deep convolutional neural networks. *Multimedia Tools and Applications*, 79(7):4499–4524, 2020. https://doi.org/10.1007/s11042-019-7684-3.

[3] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7:87–93, 2018. https://doi.org/10.1007/s13735-017-0141-z.

[4] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Indoor scene understanding with rgb-d images: bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112:133–149, 2015. https://doi.org/10.1007/s11263-014-0777-6.

[5] S. Hao, Y. Zhou, and Y. Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020. https://doi.org/10.1016/j.neucom.2019.11.118.

[6] Y. Hu, Z. Chen, and W. Lin. Rgb-d semantic segmentation: a review. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2018. https://doi.org/10.1109/ICMEW.2018.8551554.

[7] L. Li, B. Qian, J. Lian, W. Zheng, and Y. Zhou. Traffic scene segmentation based on rgb-d image and deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 19(5):1664–1669, 2017. https://doi.org/10.1109/TITS.2017.2724138.

[8] W. Li, J. Gu, Y. Dong, Y. Dong, and J. Han. Indoor scene understanding via rgb-d image segmentation employing depth-based cnn and crfs. *Multimedia Tools and Applications*, 79:35475–35489, 2020. https://doi.org/10.1007/s11042-019-07882-w.

[9] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang. Cascaded feature network for semantic segmentation of rgb-d images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1311–1319, 2017.

[10] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022. https://doi.org/10.1016/j.neucom.2022.01.005.

[11] Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung. Real-time progressive 3d semantic segmentation for indoor scenes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1089–1098. IEEE, 2019. https://doi.org/10.1109/WACV.2019.00121.

[12] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021. https://doi.org/10.1109/ICRA48506.2021.9561675.

[13] R. Velastegui, M. Tatarchenko, S. Karaoglu, and T. Gevers. Image semantic segmentation of indoor scenes: a survey. *Computer Vision and Image Understanding*, 248:104102, 2024. https://doi.org/10.1016/j.cviu.2024.104102.

[14] C. Wang, C. Wang, W. Li, and H. Wang. A brief survey on rgb-d semantic segmentation using deep learning. *Displays*, 70:102080, 2021. https://doi.org/10.1016/j.displa.2021.102080.

[15] M.-J. Yang, Y.-X. Guo, B. Zhou, and X. Tong. Indoor scene generation from a collection of semantic-segmented depth images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15203–15212, 2021.

[16] H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun, and Y. Tang. Methods and datasets on semantic segmentation: a review. *Neurocomputing*, 304:82–103, 2018. https://doi.org/10.1016/j.neucom.2018.03.037.

[17] W. Zhou, G. Xu, F. Qiang, and L. Yu. Acenet: auxiliary context-information enhancement network for rgb-d indoor scene semantic segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(2):1125–1129, 2023. https://doi.org/10.1109/TETCI.2023.3303930.

[18] L. Zhu, Z. Kang, M. Zhou, X. Yang, Z. Wang, Z. Cao, and C. Ye. Cmanet: cross-modality attention network for indoor-scene semantic segmentation. *Sensors*, 22(21):8520, 2022. https://doi.org/10.3390/s22218520.