Deep Feature Fusion Network for Lane Line Segmentation in Urban Traffic Scenes

Hoanh Nguyen
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

***Abstract*—As autonomous driving technology continues to advance at a rapid pace, the demand for precise and dependable lane detection systems has become increasingly critical. However, traditional methods often struggle with complex urban scenarios, such as crowded environments, diverse lighting conditions, unmarked lanes, curved lanes, and night-time driving. This paper presents a novel approach to lane line segmentation in urban traffic scenes with a Deep Feature Fusion Network (DFFN). The DFFN leverages the strengths of deep learning for feature extraction and fusion, aiming to enhance the accuracy and reliability of lane detection under diverse real-world conditions. To integrate multi-layer features, the DFFN employs both spatial and channel attention mechanisms in an appropriate manner. This strategy facilitates learning and predicting the relevance of each input feature during the fusion process. In addition, deformable convolution is employed in all up-sampling operations, enabling dynamic adjustment of the receptive field according to object scales and poses. The performance of DFFN is rigorously evaluated and compared with existing models, namely SCNN, ENet, and ENet-SAD, across different scenarios in the CULane dataset. Experimental results demonstrate the superior performance of DFFN across all conditions, highlighting its potential applicability in advanced driver assistance systems and autonomous driving applications.**

***Keywords*—Lane line segmentation; deep learning; convolutional neural network; spatial and channel attention**

I. Introduction

As urbanization accelerates and our reliance on transportation intensifies, the need for safer and more efficient urban traffic systems is more pressing than ever. Among the numerous challenges in developing intelligent transportation systems, accurate lane line segmentation is a vital task for the functionality of autonomous driving and advanced driver-assistance systems (ADAS). By properly recognizing and predicting lane lines, these systems can better ensure the safety and efficiency of road traffic. Within the domain of computer vision and image processing, a plethora of methods have been proposed to address lane line segmentation. Traditional methods, like edge detection and Hough transform, provide some utility, but their performance can be significantly hindered under complex conditions such as variable lighting, weather, and diverse road markings. With the rise of deep learning, Convolutional Neural Networks (CNNs) have shown superior performance in various tasks including lane line segmentation [1], [2]. In recent years, a range of techniques for lane line segmentation employing Convolutional Neural Networks (CNNs) have been devised. Study [3] presents a robust method for lane detection in continuous driving scenarios, leveraging the power of deep neural networks. The authors introduced a novel two-stage framework that first generates lane line proposals using a pixel-wise prediction model, and then refines these proposals through a sequential prediction model, leveraging temporal information between frames. Their method demonstrated impressive robustness in handling various complex scenarios and achieved notable performance on multiple benchmark datasets. Philion [4] proposed a novel method to tackle the "long tail" problem in lane detection - the issue of detecting rare or unusual lane configurations. The approach uses a sequential prediction network that dynamically generates waypoints, thereby allowing it to adapt to a wide variety of lane shapes and configurations. In their study, Qin, Wang, and Li (2020) [5] introduced a structure-aware deep lane detection algorithm. The algorithm focuses on improving the speed and efficiency of lane detection by incorporating prior structural knowledge into a novel deep learning framework. Recently, Yoo et al. [6] proposed an end-to-end lane marker detection algorithm using a row-wise classification approach in their research. Their method transforms the challenging lane detection problem into a simpler row-wise classification task, improving both speed and accuracy of detection. Another approach [7] is to utilize a fully convolutional neural network with a novel instance segmentation head to simultaneously detect and separate different lane lines. In recent work, Abualsaud et al. [8] introduced LaneAF, a robust multi-lane detection method based on the concept of affinity fields. The proposed approach uses the affinity fields to encode relational information between different parts of the lane lines, enhancing the detection accuracy in challenging situations like close, parallel, and curvy lanes. Wang, Ren, and Qiu [9] introduced LaneNet, a real-time lane detection network designed for autonomous driving applications. LaneNet utilizes a two-branch neural network that simultaneously performs semantic segmentation for pixel-wise lane detection and instance segmentation for distinguishing between individual lane lines. In [10], Pan et al. introduced a novel concept of Spatial Convolutional Neural Networks (SCNN) that extends traditional CNNs by performing convolutions in the spatial domain. This novel SCNN framework, which treats spatial information as a type of deep information, was shown to be particularly effective in traffic scene understanding tasks, including lane line detection. Based on SCNN, Zheng et al. [11] proposed a Recurrent Feature-Shift Aggregator (ReSA) for lane detection tasks. The ReSA model uses a novel recurrent structure to shift and aggregate deep features, effectively capturing the spatial dependencies of lane pixels and thereby improving lane detection performance. Hou et al. [13] introduced a self-attention distillation strategy for developing lightweight lane

detection CNNs. The method involves training a smaller student network to mimic the attention maps of a larger, pre-trained teacher network, thereby improving the efficiency and performance of the student network. More recently, Vu et al. [14] proposed HybridNets, an end-to-end perception network for autonomous driving. HybridNets, combining multiple sub-networks tailored to different perception tasks, provides a unified architecture that can simultaneously perform various tasks, including lane line detection, while sharing learned representations. In addition to structures specifically designed for the task of lane line segmentation, some proposed methods use popular networks for general semantic segmentation such as Fully Convolutional Networks (FCN) [14], U-Net [15], SegNet [16], DeepLab v3+ [17], for the task of lane line segmentation, which also yield promising results and achieve significant outcomes.

Although the above methods show promise in lane line segmentation tasks, challenges remain due to the intricate nature of urban scenes that include varying lanes, unpredictable surrounding environments, and complicated traffic scenarios. To address these challenges, this paper presents a Deep Feature Fusion Network (DFFN) for lane line segmentation in urban traffic scenes. The core idea of the proposed method is to leverage the strength of deep learning and feature fusion to extract and combine multi-level and multi-scale features from the input images. This approach not only enhances the robustness of the network against complex conditions but also significantly improves the segmentation performance by effectively capturing both the local detailed information and the global contextual information of lane lines. The effectiveness of the proposed model has been verified through experiments on the CULane dataset.

The rest of this paper is organized as follows: Section II presents the detailed methodology of the proposed model, including the architectural design, key components, and training procedure. Section III describes the CULane dataset and the experimental setup, followed by a comprehensive analysis of the experimental results. Section IV summarizes the key contributions and highlighting the significance of the proposed model in improving the functionality and safety of autonomous driving systems and advanced driver assistance systems.

II. METHODOLOGY

This section elaborates on the proposed DFFN structure designed for lane line segmentation in urban traffic scenes. DFFN is based on the DLA structure [18] used for the semantic segmentation task. Therefore, this section will first provide a summary of the DLA network structure, followed by a detailed explanation of the proposed modifications designed to enhance the DLA model specifically for the lane line segmentation problem.

A. DLA Network for Semantic Segmentation

Deep Layer Aggregation is a powerful structure that has seen successful applications across a variety of computer vision tasks, including semantic segmentation. The design of DLA is based on the observation that semantic segmentation requires not only high-level semantic information but also low-level detailed information. The main aim of the DLA architecture is to effectively aggregate multi-scale and multi-level features to generate rich and detailed feature maps that are beneficial for tasks like semantic segmentation. DLA consists of two major components: a hierarchy of basic blocks and an aggregation mechanism, as shown in Fig. 1(a).

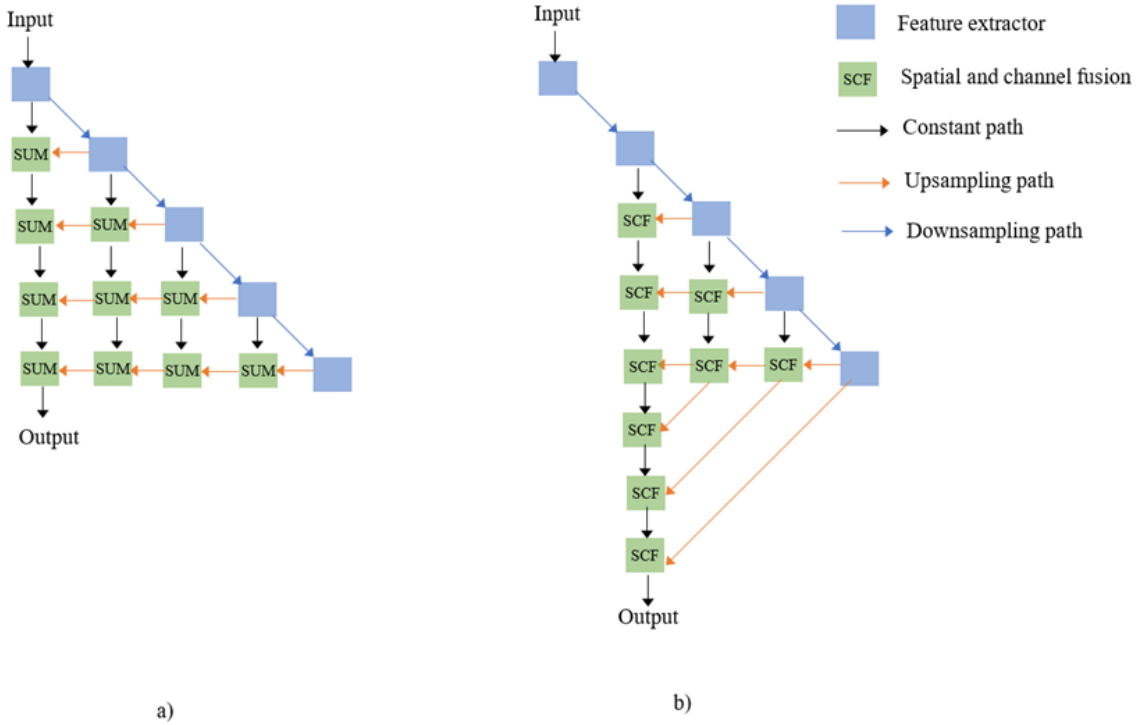

Fig. 1. The structure of original DLA (a) and the proposed DFFN (b).

1) Hierarchy of basic blocks: DLA adopts a hierarchical structure similar to typical convolutional networks such as ResNet [19] or ResNeXT [20], but with each level consisting of basic blocks, with each block being a small network of its own. Each basic block within the hierarchy operates at a different resolution, and the block output is a feature map of the corresponding resolution. Lower-level blocks capture fine-grained features, while higher-level blocks capture coarser but more abstract features.

2) Aggregation mechanism: The uniqueness of DLA lies in its aggregation mechanism. Traditional convolutional networks only use features from the highest level for prediction, which could result in a loss of detailed spatial information. DLA, however, introduces an aggregation mechanism that propagates the information from higher layers to lower layers, in a top-down manner. This aggregation allows high-level semantic features to be combined with low-level spatial features. The process begins with the highest level, where features are first processed by a 1×1 convolution to reduce the channel dimension. Then, these features are upsampled and summed with the corresponding lower-level features. The combined features are then processed by another 1×1 convolution before being passed to the next lower level. The aggregation mechanism allows DLA to generate rich feature maps that contain both high-level semantic information and low-level detailed information. This feature is particularly beneficial for semantic segmentation, which requires a good understanding of both the object (high-level) and the exact boundary (low-level) of each semantic class.

B. Deep Feature Fusion Network with Spatial and Channel Fusion

Although DLA has achieved some success in semantic segmentation tasks, its performance in lane line segmentation in urban traffic scenes is greatly limited. There are several reasons to explain this. Firstly, the proportion of lane lines usually occupies a relatively small ratio in the image, and sometimes lane lines are not clearly visible. This severely restricts the accuracy of pixel-level segmentation of lane lines. Secondly, in complex environments where lane changes, changing lighting conditions, or irregular lane shapes frequently occur, the feature fusion scheme in DLA is easily affected by background noise. Inspired by attention mechanism [21], which employs channel and spatial self-attention for adaptive feature refinement to enhance the performance of convolutional networks in tasks like image classification, image captioning, and object detection, this paper designs the DFFN based on the DLA architecture for efficient lane line segmentation in urban traffic scenes. Fig. 1(b) illustrates the detailed structure of the proposed DFFN. It judiciously employs both spatial and channel attention mechanisms to learn and anticipate the significance of each input feature during the fusion process. Consequently, it amplifies lane line features from both spatial and channel dimensions, extracting effective lane line characteristics even in challenging environments. Specifically, DFFN utilizes ResNet-34 [19] as its backbone to create an optimal balance between precision and processing speed. It deviates from the traditional DLA by integrating more skip connections between low-level and high-level features, resembling the operational structure of the Feature Pyramid Network [22]. Moreover, DFFN replaces the convolution layers in all up-sampling modules with deformable convolution, allowing for dynamic adjustments of the receptive field in accordance with object scales and orientations. This transformation not only offers flexibility but also helps mitigate alignment issues. In addition, each linear aggregation node in the original DLA structure is replaced by the spatial and channel fusion node (SCF), which is designed to compute spatial and channel attention based on the relation of the input feature maps. The next subsection will elaborate on the spatial and channel fusion design.

1) Spatial and channel fusion: The spatial and channel fusion is applied on two different input feature maps, I_S and I_L, where I_S is the shallower, higher resolution feature map and I_L is the deeper, lower resolution feature map, as shown in Fig. 2. Since I_S contains richer spatial information, this paper applies spatial attention operation on this feature map to enhance its spatial information. The spatial attention operation includes two 3×3 convolution layers followed by sigmoid activation. Suppose $I_S \in \mathbb{R}^{W\times H\times C}$, the output of the spatial attention operation I'_S is calculated as follow:

$$I'_S = \sigma(h(I_S)) \quad (1)$$

where $h(.)$ is the convolution operation, and σ is the sigmoid function.

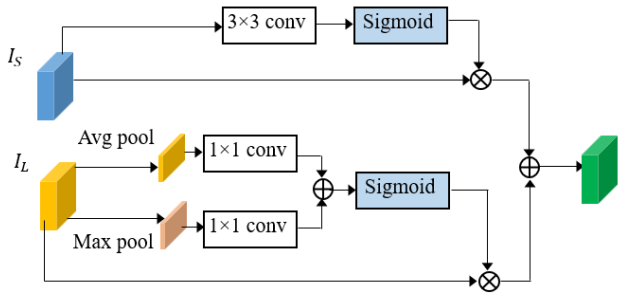

Fig. 2. Spatial and channel fusion.

On the other hand, as I_L have richer semantic representations, this paper applies channel attention operation on this feature map to improve its channel features. The channel attention operation first applies average pooling and max pooling to generate intermediate feature maps. Then, two 1×1 convolution layers are applied in parallel to further transform these intermediate feature maps. Finally, the sigmoid activation function is used after summing the intermediate maps to generate the rich channel semantic maps. Suppose $I_L \in \mathbb{R}^{W\times H\times C}$, then the output of the channel attention operation I'_L is calculated as follow:

$$I'_L = \sigma(g(Avg(I_L)) + g(Max(I_L))) \quad (2)$$

After computing spatial and channel attention based on the relation of the input feature maps, this paper employs element-wise multiplication and summation to generate final enhanced feature map as follow:

$$I_O = I'_S \odot I_S + I_L \odot I'_L \quad (3)$$

Since the spatial and channel fusion module employs simple non-linear operation, it introduces negligible computation overhead.

III. RESULTS

A. Dataset and Metrics

This paper employs the CULane dataset [10] to evaluate the proposed model. The CULane dataset has been utilized in various studies related to autonomous driving and advanced driver assistance systems. It's especially popular for tasks such as lane detection, semantic segmentation, and traffic scene understanding. The CULane dataset is quite large, containing around 55,000 images, and covering various scenarios with different traffic, lighting, and weather conditions. It consists of images from urban streets, highways, and rural areas captured at different times of the day. It also includes challenging driving scenarios like night driving, shadows, dazzling, and rainy or foggy conditions, thus offering a comprehensive dataset for robust model training. Each image in the CULane dataset is carefully annotated with high-quality pixel-level annotations of lane lines, including markings for straight lanes, curved lanes, and parallel lanes. This detailed annotation serves as an excellent training ground for lane segmentation models. It is worth noting that each image also contains corresponding binary lane segmentation maps, which are quite useful for model training and evaluation. The dataset is split into distinct training and testing sets, providing a reliable platform for both the development and evaluation of models. The training set contains around 88,880 images, while the test set contains approximately 34,680 images, spread across 9 different categories representing a range of driving conditions, as shown in Table I and Fig. 3. This paper carefully screened 40,000 annotated images containing lane lines in the dataset and used 70% of the filtered dataset for training. As in [10], this paper uses $F1$-measure as metric for evaluating the proposed model.

Fig. 3. Some examples for different scenarios.

TABLE I. PROPORTION OF EACH CATEGORY IN THE CULANE DATASET

Category	Proportion (%)	Resolution
Normal	27.7	590×1640
Crowded	23.4	
Dazzle light	1.4	
Shadow	2.7	
No line	11.7	
Arrow	2.6	
Curve	1.2	
Night	20.3	
Crossroad	9.0	

B. Experimental Results

This paper compared the performance of the proposed method against established models including SCNN [10], ENet [23], and ENet-SAD [12]. All experiments were conducted across eight distinct categories of the CULane testing set, evaluated based on F1-measure. The results are shown in Table II. In the Normal condition, DFFN demonstrated superior performance with an F1 score of 70.25%, compared to SCNN (60.12%), ENet (65.62%), and ENet-SAD (67.72%). Under Crowded circumstances, the robustness of the DFFN model was notable, achieving an F1 score of 58.71%, outperforming SCNN (45.38%), ENet (55.46), and ENet-SAD (55.81). In the Dazzle light and Shadow scenarios, DFFN continued to excel, achieving F1 scores of 53.54% and 55.62% respectively, surpassing the scores of SCNN, ENet, and ENet-SAD. For the No line and Arrow conditions, DFFN maintained high performance levels, demonstrating impressive lane recognition capability in comparison to other models, as evidenced by the F1 scores. In the Curve category, DFFN achieved an F1 score of 58.80%, demonstrating superior performance in identifying and tracking curved lanes. Lastly, in the Night condition, DFFN upheld its strong performance, with an F1 score of 58.62%, outperforming the compared models in low-light conditions. These experimental results underscore the effectiveness and robustness of the proposed DFFN method across varied traffic scenarios and lighting conditions. The consistently high F1 scores, in comparison to other established models, suggest promising potential for DFFN in real-world applications, such as autonomous driving and advanced driver assistance systems.

Fig. 4 provides a detailed visual comparison of the performance of the proposed DFFN, SCNN, and ENet on the CULane testing images. The first column displays the original image, providing the actual scene context from the CULane testing set. The second column shows the ground truth, representing the ideal output that the models should aim to replicate. These images provide a benchmark against which the model outputs are evaluated. The third column presents the results of the proposed DFFN model. An initial visual comparison between these outputs and the ground truth may suggest the effectiveness of the DFFN in accurately segmenting lane lines under different traffic scenarios. The fourth column illustrates the output from the SCNN model. By comparing these images with the ground truth and the DFFN outputs, we can assess the performance of the SCNN in relation to both the ideal output and the proposed model. The final column depicts the results from the ENet model. Again, a comparison between these images, the ground truth, and the other model outputs helps evaluate the performance of the ENet model in various traffic conditions. A detailed examination of Fig. 4 would provide insights into the areas where the proposed DFFN outperforms or underperforms compared to the SCNN and ENet models. For example, we might observe that the DFFN model performs particularly well in crowded scenarios or shadow conditions, offering more accurate and robust lane segmentation than the comparative models. However, the DFFN might be sensitive to noise and outliers in the input data. This could result in misclassification or incomplete segmentation of lane lines, affecting the overall accuracy and reliability of the model's outputs. Addressing the

sensitivity to noise and outliers is an important challenge in lane line segmentation with deep learning models like the DFFN. Techniques such as data augmentation, robust feature extraction, and outlier detection can be explored to improve the model's resilience to noisy input data and enhance its accuracy and reliability.

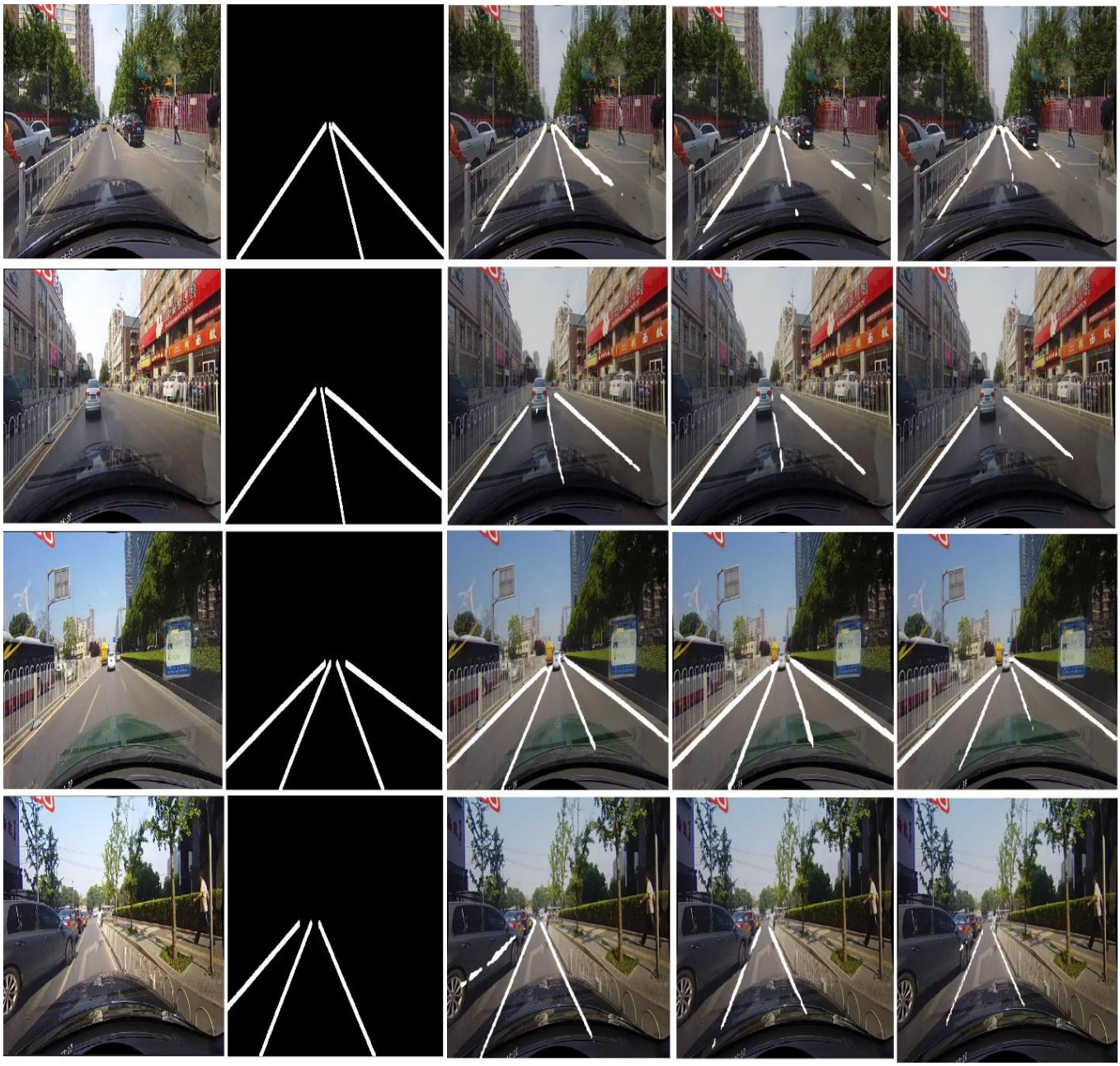

Fig. 4. Visualization of experimental results in CULane dataset of the proposed model and SCNN, ENet.

TABLE II. F1-MEASURE (%) OF DIFFERENT APPROACHES ON THE CULANE TESTING SET

Method	Category							
	Normal	Crowded	Dazzle light	Shadow	No line	Arrow	Curve	Night
SCNN [10]	60.12	45.38	37.52	41.44	36.34	45.31	44.44	41.20
ENet [23]	65.62	55.46	50.21	54.49	35.82	58.11	56.43	49.39
ENet-SAD [12]	67.72	55.81	52.91	54.51	39.06	56.94	57.91	54.12
Proposed model	70.25	58.71	53.54	55.62	39.86	59.17	58.80	58.62

IV. CONCLUSION

This paper introduces the Deep Feature Fusion Network (DFFN), a novel approach for lane line segmentation in complex urban traffic scenes. Based on the DLA structure, the DFFN integrates more skip connections between low-level and high-level features. In addition, each linear aggregation node in the original DLA structure is replaced by the spatial and channel fusion node to learn and predict the importance of each input feature during the fusing process. The DFFN has demonstrated its robustness in challenging scenarios, including crowded environments, varying lighting conditions, unmarked lanes, and curved paths, outperforming established models consistently. These results highlight the potential of the DFFN model in improving the functionality and safety of autonomous driving systems and advanced driver assistance systems. Despite its current performance, there is always room for improvement and optimization. Future work could focus on further enhancing the DFFN's ability to adapt to diverse environmental conditions and refining the model's capability to handle more complex and unusual lane line patterns, as well as addressing the sensitivity to noise and outliers.

REFERENCES

[1] Nisa, Syed Qamrun, and Amelia Ritahani Ismail. "Dual U-Net with Resnet Encoder for Segmentation of Medical Images." *International Journal of Advanced Computer Science and Applications* 13, no. 12 (2022).

[2] Marcellino, & Cenggoro, Tjeng Wawan & Pardamean, Bens. (2022). UNET++ with Scale Pyramid for Crowd Counting. ICIC Express Letters. 16. 75-82. 10.24507/icicel.16.01.75.

[3] Zou, Qin, Hanwen Jiang, Qiyu Dai, Yuanhao Yue, Long Chen, and Qian Wang. "Robust lane detection from continuous driving scenes using deep neural networks." *IEEE transactions on vehicular technology* 69, no. 1 (2019): 41-54.

[4] Philion, Jonah. "Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11582-11591. 2019.

[5] Qin, Zequn, Huanyu Wang, and Xi Li. "Ultra fast structure-aware deep lane detection." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 276-291. Springer International Publishing, 2020.

[6] Yoo, Seungwoo, Hee Seok Lee, Heesoo Myeong, Sungrack Yun, Hyoungwoo Park, Janghoon Cho, and Duck Hoon Kim. "End-to-end lane marker detection via row-wise classification." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1006-1007. 2020.

[7] Neven, Davy, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. "Towards end-to-end lane detection: an instance segmentation approach." In *2018 IEEE intelligent vehicles symposium (IV)*, pp. 286-291. IEEE, 2018.

[8] Abualsaud, Hala, Sean Liu, David B. Lu, Kenny Situ, Akshay Rangesh, and Mohan M. Trivedi. "Laneaf: Robust multi-lane detection with affinity fields." *IEEE Robotics and Automation Letters* 6, no. 4 (2021): 7477-7484.

[9] Wang, Ze, Weiqiang Ren, and Qiang Qiu. "Lanenet: Real-time lane detection networks for autonomous driving." *arXiv preprint arXiv:1807.01726* (2018).

[10] Pan, Xingang, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Spatial as deep: Spatial cnn for traffic scene understanding." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. 2018.

[11] Zheng, T., Fang, H., Zhang, Y., Tang, W., Yang, Z., Liu, H. and Cai, D., 2021, May. Resa: Recurrent feature-shift aggregator for lane detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 4, pp. 3547-3554).

[12] Hou, Y., Ma, Z., Liu, C. and Loy, C.C., 2019. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1013-1021).

[13] Vu, Dat, Bao Ngo, and Hung Phan. "Hybridnets: End-to-end perception network." *arXiv preprint arXiv:2203.09035* (2022).

[14] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440. 2015.

[15] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234-241. Springer International Publishing, 2015.

[16] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 12 (2017): 2481-2495.

[17] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801-818. 2018.

[18] Yu, Fisher, Dequan Wang, Evan Shelhamer, and Trevor Darrell. "Deep layer aggregation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2403-2412. 2018.

[19] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[20] Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492-1500. 2017.

[21] Woo, Sanghyun, Jongchan Park, Joon-Young Lee, and In So Kweon. "Cbam: Convolutional block attention module." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3-19. 2018.

[22] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125. 2017.

[23] Paszke, Adam, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. "Enet: A deep neural network architecture for real-time semantic segmentation." *arXiv preprint arXiv:1606.02147* (2016).