

Steering Angle-Guided Multimodal Fusion Lane Detection for Autonomous Driving

Yan Gong¹, Xinyu Zhang¹, *Member, IEEE*, Jianli Lu², Xinmin Jiang, Zichen Wang, Hao Liu, Zhiwei Li³, Li Wang⁴, Qingshan Yang⁴, and Xingang Wu

Abstract—Lane detection is a critical part of autonomous driving technology. When difficult situations are encountered (i.e., adverse light, severe occlusion), the lane detection task is still challenging. However, previous methods strongly depend on the extracted image features and ignore other features. It is necessary to consider the information from other modalities to assist the model for lane detection, especially in the task of curved lane detection. In this paper, considering that the vehicle steering angle is closely related to the visual feature of lane lines, we propose a novel model named Image-Angle Fusion Network (IAFNet) to solve the lane detection problem by fusing vehicle steering angle features with image features. To make the steering angle features better match the image features, we use the tensor outer product to extend the dimensionality of the steering angle information. A lightweight Image-Angle cross-attention module (LIA-CAM) is proposed to learn the implicit relationship between steering angles and visual features of lane lines, aimed at improving the performance of our model in difficult situations. To guide the network to retain the correct steering angle information, we introduced regression prediction loss of steering angle. Besides, we also released a new dataset based on the Udacity dataset: ImageAngle-Udacity (IA-Udacity) dataset. Extensive experiments on the IA-Udacity dataset show that our method outperforms the current state-of-the-art methods showing both higher efficiency and accuracy. Code and data are available on <https://github.com/gongyan1/LIA-CAM>.

Index Terms—Lane detection, multimodal fusion, steering angle, deep learning, convolutional neural network.

I. INTRODUCTION

AUTOMATIC lane detection is an important part of autonomous driving and the Advanced Driver Assistance System (ADAS), aims to detect the precise location and type information about lanes [1], [2], [3], [4]. This information is crucial for self-position constraint and path planning of vehicles [5], [6].

With the development of deep learning in recent years [7], many convolutional neural networks (CNNs) [8], [9], [10], [11] based methods for lane detection have been proposed. Benefiting from the powerful feature extraction and representation capabilities of CNNs, they both achieve promising performance. LaneNet [12] detects locations of lanes through a binary segmentation branch. To leverage the strong shape priors of lanes, SCNN [13] and RESA [14] focus on improving the model's ability to capture spatial relationships of pixels across rows and columns, towards learning more global features. Inspired by the anchor-based object detection methods [10], Line-CNN [15], LaneATT [16], SGNet [17] adopt line anchors and follow a top-down roadmap to regress pre-defined anchors into various shapes of lane lines. All methods solely use pure visual information (i.e., images taken by cameras mounted on vehicles) for lane detection [18], [19], [20]. The detection performance is highly dependent on whether CNNs can extract highly discriminative features. However, there are some shortcomings. First, they ignore the strong shape priors of lanes and high correlations between lanes in one image. Second, CNNs often have difficulty learning accurate features in the presence of no visual clues (e.g., lanes occluded by vehicles or dashed lanes) or in extreme lighting and weather conditions, resulting in poor detection performance, as is shown in Fig. 1. Only CNNs are used to extract the feature information of lane lines, which is not conducive to the task of curved lane line detection in complex environments.

In complex driving environments, a single sensor information is not enough to effectively deal with changes in the scene [21]. Therefore, the complementary of different modal information will be more important in the task of lane line detection. There are many works on multimodal fusion nowadays. TFN [22] computes the correlations between elements of different modalities via the tensor outer product.

Received 27 March 2023; revised 16 April 2024 and 29 July 2024; accepted 7 November 2024. Date of publication 23 December 2024; date of current version 4 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62273198 and Grant 52221005, in part by Beijing Natural Science Foundation Program under Grant L241017, and in part by the National Research Foundation Singapore under its AI Singapore Program under Award AISG2-GC-2023-007. The Associate Editor for this article was C. Guo. (*Corresponding author: Xinyu Zhang.*)

Yan Gong is with the State Key Laboratory of Intelligent Green Vehicle and Mobility, School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China, and also with the Autonomous Driving Department of X Division, JD Logistics, Beijing 101111, China (e-mail: gongyan2020@foxmail.com).

Xinyu Zhang is with the State Key Laboratory of Intelligent Green Vehicle and Mobility, School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China, and also with the Electrical and Computer Engineering Department, National University of Singapore, Singapore 117583 (e-mail: xyzhang@tsinghua.edu.cn).

Jianli Lu, Xinmin Jiang, Zhiwei Li, Qingshan Yang, and Xingang Wu are with the State Key Laboratory of Intelligent Green Vehicle and Mobility, School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China (e-mail: lujianli364@163.com; xinminj2001@gmail.com; 2022500066@buct.edu.cn; qingshan.yang@hirain.com; wuxingang@tsinghua.edu.cn).

Zichen Wang and Hao Liu are with the Autonomous Driving Department of X Division, JD Logistics, Beijing 101111, China (e-mail: wangzichen@jd.com; liuhao163@jd.com).

Li Wang is with the School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China (e-mail: wangli_bit@bit.edu.cn).

Digital Object Identifier 10.1109/TITS.2024.3507536

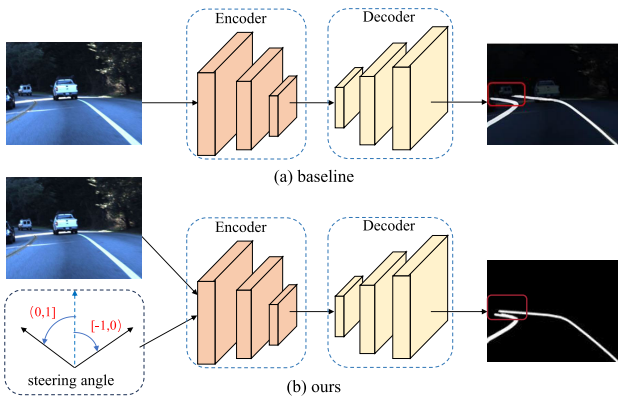


Fig. 1. Flowchart of different networks used for lane detection. The baseline (a) is the flowchart without steering angle, and ours (b) is the flowchart with fused steering angle features. It is worth noting that we normalize the steering angle to $[-1.0, 1.0]$, where a negative number represents a right turn and a positive number represents a left turn.

LMF [23] uses low-rank matrix factorization to reduce the amount of calculation in the above process of calculating tensor outer products. Inspired by Transformer, MuIT [24] and MBT [25] use a cross-attention mechanism, transforming the data of one modality into query and the data of another modality into key and value, and adaptively fuse multi-modal information by modeling relationships between query and key. Facing the problem of detecting curved lane lines in an adverse lighting environment, the above proposed multimodal fusion method cannot extract the steering angle features in curved lane lines and cannot effectively assist the network model to extract the lane line features in images.

For vehicles, in addition to the camera, they are also equipped with a variety of sensors to obtain different information, such as steering angle sensors, LiDAR, etc [26]. For example, TransFusion [27] uses LiDAR and image fusion to perform 3D object detection [28], [29], [30]. These sensors reflect the driving state of the vehicle and the surrounding environment. Among them, the steering angle sensor reflects the steering angle of the front wheel. However, no method has been previously proposed for the multimodal fusion of steering angles and images. In the case of adverse lighting and curved lane lines, we believe that steering angle information is important to assist the network model in accurately segmenting lane lines. The steering angle has a great relationship with the visual characteristics of the current lane. When the car is driving straight, straight lanes and the front wheel will form a nearly parallel spatial relationship. When the car turns left, straight lanes form an angle with the front wheel, which is expressed in the form of the steering angle. For curved lanes, there is also an implicit spatial relationship between lanes and the steering angle of the vehicle. The process of vehicle deviation or lane change will produce a switch of the lane from where the vehicle is located, and a switch of the lane line to the left/right line will occur [31], [32]. Similarly, lane line transitions involve vehicle steering angle information. In summary, the vehicle steering angle can provide good prior knowledge for the detection of lanes, improving the performance of lane detection in difficult situations such as

no visual clues or extreme lighting and weather, as shown in Fig. 1. After adding the steering angle feature to the network, the steering angle information will assist the network to learn the image features better, and the image features can also guide the steering angle information. The two complement each other, thus improving the performance of the network.

In this paper, we propose a multimodal lane detection model that fuses the steering angle of a vehicle with the images captured by the camera on the vehicle, and uses the fused features for lane detection. In this way, our model can use both image features and steering angle features to detect lane lines more effectively. In general, we first consider the lane detection problem as a semantic segmentation problem, i.e., detecting the specific location of lane lines by classifying the images pixel by pixel. We adopt the Encoder-Decoder network structure to learn and segment the image features to solve the problem of curved lane detection in adverse lighting environments. In this process, the image features are attentively fused with steering angle features of equal size several times. We adopt a lightweight image-angle cross-attention module (LIA-CAM) to fuse the information of these two modalities. We leverage LIA-CAM for multimodal fusion to learn the a priori knowledge of the relationship between steering angle and lane line visual features to obtain better lane line segmentation performance. To obtain steering angle features of equal size to the image features, we apply multiple tensor outer product operations to extend the scalar steering angle to a dimension consistent with the image features. Besides, our model finally performs the steering angle regression prediction task, which acts as a supervisory signal for the network at the same time as the lane segmentation loss. Such an operation motivates the network to maintain the correct steering angle characteristics from start to finish. To balance the losses of these two tasks and reduce the performance degradation of the two losses, we adopt the method in [33] to adaptively assign the losses of the split task to the losses of the regression task.

Our contributions can be summarized in the following three points:

1) We are the first work to introduce steering angle information into lane detection, which adopts LIA-CAM to fuse steering angle features with image features, allowing the network to exploit the implicit relationship between the two to assist lane detection in adverse lighting environments.

2) We propose the Image-Angle Fusion Network (IAFNet), an encoder-decoder structure network to detect curved lane lines by performing semantic segmentation. Furthermore, we introduce an angle regression loss to induce the network to retain the correct value through the network.

3) We release a new dataset based on the Udacity dataset: ImageAngle-Udacity (IA-Udacity) [34], which contains part of the images in Udacity. We annotate each image with lane line masks and normalize the corresponding steering angle for each image.

4) We conduct intensive experiments on the above dataset. As a result, our network obtains performances that exceed multiple previous state-of-the-art methods, reaching 78.41% IoU and 87.99% F1-score respectively.

II. RELATED WORK

A. Lane Detection Methods

Current deep-learning-based lane detection methods can be broadly classified into several categories. In this section, we will elaborate on these methods separately.

1) *Segmentation-Based Methods*: Segmentation-based methods treat the lane detection problem as a pixel-wise classification problem and perform lane recognition by classifying each pixel in the image into lane class or background class. LaneNet [35] casts the lane detection problem as an instance segmentation problem. It mainly consists of a lane segmentation branch and a lane embedding branch to perform binary segmentation and disentangle the segmented lane pixels into different lane instances respectively. SCNN [13] proposes slice-by-slice convolutions to capture spatial relationships of pixels across rows and columns, which is suitable for learning objects with strong shape priors but weak appearance coherences (e.g., lanes). Similar to SCNN [13], RESA [14] proposes a recurrent feature-shift aggregator to shift sliced feature maps recurrently, enabling global information gathering. ENet-SAD [36] presents a self-attention distillation approach, to leverage the learned attention maps to encode rich contextual information. Inspired by segmentation-based methods, we combine steering angle information with image feature information and adopt an attention mechanism for feature learning.

2) *Detection-Based Methods*: These methods can be divided into line-anchor-based methods and row-anchor-based methods. The former detects lanes by regressing the offsets between sampled points on line anchors and predefined anchors. Line-CNN [15] is a pioneering work in line-anchor-based methods, which emits a series of line anchors from the boundary of a particular direction of the image. LaneATT [16] proposes a novel anchor-based feature pooling and attention mechanism to aggregate more global information. Row-anchor-based methods, on the other hand, predict the grids in predefined rows in the image where lanes are likely to exist. UFLD [37] firstly treats the process of lane detection as a row-based selecting problem using global features. Besides, a structural loss is also proposed to explicitly model the structure of lanes. CondLaneNet [35] introduces a conditional lane detection strategy based on conditional convolution and row-wise formulation. It first detects the lane instances and then dynamically predicts the line shape for each instance. Other works formulate lane detection as a key-point estimation and correlation problem to describe the shapes of lanes with more flexibility. They gradually group adjacent key points belonging to the same lane in a point-by-point manner [38] or a parallel manner [15]. Different from the above methods, we did not adopt the way of lane detection through anchor points but designed a multimodal fusion lane detection network by using segmentation.

3) *Keypoints-Based Methods*: Some works formulate lane detection as a key-point estimation and correlation problem. FOLOLane [38] models local patterns by two separate heads, where the first one predicts the existence of key points and the second one refines the locations and correlates key points of

the same lane. GANet [39] predicts each key point by regressing it with the start of the lane to which it is subordinate. Furthermore, a lane-aware feature aggregator is also proposed to adaptively capture relations between adjacent key points to provide local information to the global association. The detection algorithm based on key points does not require post-processing, so the ability to describe the curve is limited by the anchor, resulting in the loss of feature information. Through the multimodal fusion of steering angle feature information and image feature information, our model can better ensure the integrity of lane line feature information.

B. Multimodal Information Fusion Methods

There is heterogeneity between the data of different modalities, and it is worth considering how to fuse them in a way that can take full advantage of their complementarity while reducing fusion redundancy for the purpose of improving the performance of the model. Overall, current multimodal fusion methods based on deep learning can be broadly classified into three categories: early fusion, deep fusion, and late fusion, depending on the stage at which fusion occurs, i.e., data-level, feature-level, or proposal-level. TFN [22], which belongs to the early fusion, proposes a model named Tensor Fusion Network, which calculates the correlation between different modal elements through the outer product. LMF [40] reduces the exponential time complexity in TFN to linear by decomposing the tensor and weights in parallel to perform multimodal fusion using mode-specific low-order factors. Previous tensor-based multimodal fusion techniques only consider bilinear or trilinear pooling, which fail to unleash the expressive power of multilinear fusion and ignore the complex local intercorrelations, thus a polynomial tensor pooling is proposed in PTP [41] to integrate multimodal features by considering high-order moments, followed by a tensorized fully connected layer.

After the emergence of the attention mechanism, several multimodal fusion methods based on the attention mechanism have been proposed. It is worth noting that the following multimodal fusion methods based on attention mechanisms all belong to deep fusion. MFN [42] propose a Delta-memory Attention Network and Multi-view Gated Memory to model multi-modal interactions. Specifically, the memory aims to preserve the multimodal information of the previous moment, where the gating mechanism is responsible for filtering the information and the attention mechanism is responsible for assigning different weights. With the success of Transformer [43] in the field of natural language processing, this structure has been introduced into the field of computer vision, such as classification [44], [45], object detection [46], [47], and segmentation [48], [49], [50]. Considering the natural ability of the cross-attention mechanism in Transformer [43] to model relationships between two sets of feature maps, there are many current works using it to achieve multimodal information fusion. Mult [24] introduces a directional pairwise cross-modal attention, which attends to interactions between multimodal sequences across distinct time steps and latently adapts streams from one modality to

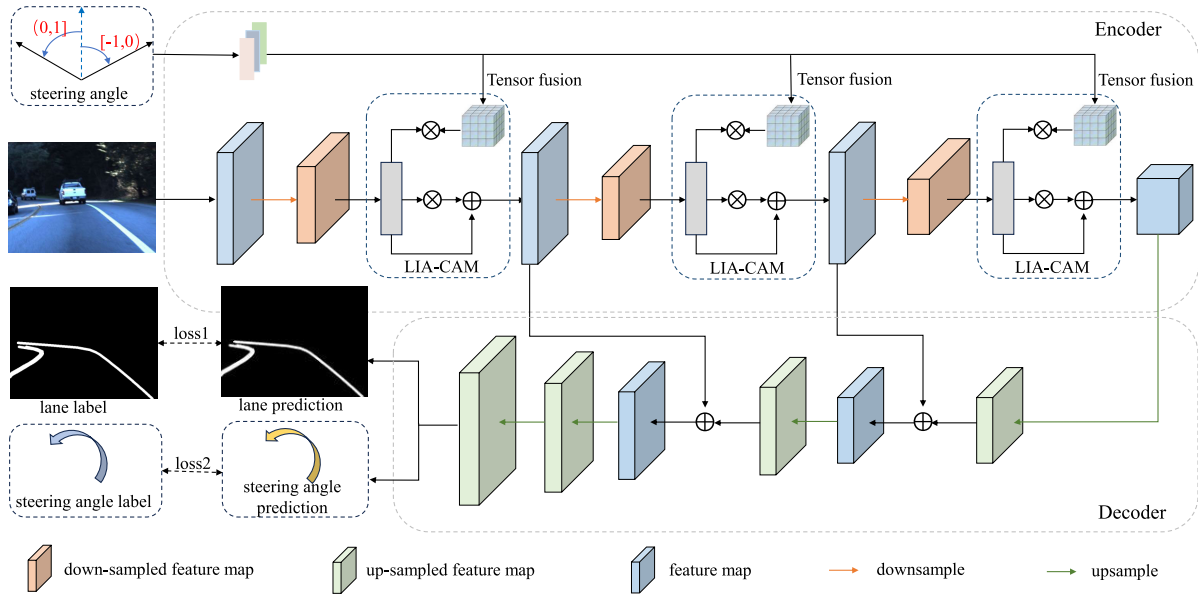


Fig. 2. The overall network architecture of IAFNet. The network structure contains two input branches, e.g., an image branch and a steering angle branch. The steering angle branch is used in the encoder to fuse features with the latter three encoder feature maps to guide the network learning. The decoder recovers the details of the target and the corresponding spatial dimension through the upsampling network layer. Finally, the network predicts the segmentation maps and steering angle results. The structure diagrams of Tensor Fusion and LIA-CAM are introduced in detail in Fig. 3 and Fig. 4, respectively. The loss functions of loss1 and loss2 are described in detail in Section III-C.

another. Neural Machine Translation with Universal Visual Representation [51] introduces an aggregation module between the encoder and decoder of a naive Transformer, which utilizes a cross-attention mechanism to fuse the corresponding image features to assist the machine translation. In MBT [25], it introduces a shared token between two Transformers, thus the token becomes the communication bottleneck of different modalities to reduce the cost of computing attention weights. In the field of autonomous driving perception, multimodal sensor information fusion has been a very critical task. TransFusion [27] proposes a spatially modulated cross-attention mechanism to attentively associate and fuse the object queries (extracted from LiDAR features) with the image features, leading to a robust and effective fusion strategy. Maff-net [52] proposes a module named dense attention fusion that fuses pseudo images from multiple branches. EPNet [53] introduces a deep LIDAR-Image fusion strategy that estimates the importance of corresponding image features to reduce the noisy influence. For the first time, we fuse steering angle information and image feature information in a multimodal strategy to guide network training for lane detection.

III. METHODS

A. Overall Structure

The overall structure of IAFNet is shown in the Fig. 2. IAFNet has two input branches, one branch is the image and the other branch is the current steering angle of the vehicle in the image. In the encoder stage, the steering angle features are extended to the image feature dimension size by the tensor outer product. The LIA-CAM is applied to the steering angle feature map and the image feature map. Different from the single-layer fusion in other methods, we perform multi-level fusion here, so that the image features and steering angle information are fully integrated. Through the attentive

calculation between the two, the information of the vehicle corner will be fused with the image features to assist the network in segmenting the lane lines accurately. In the decoder stage, we use the upsampling method of deconvolution to recover the feature map size. And the feature maps of the two stages are added with the corresponding feature maps of the encoder. This operation can preserve multi-scale information, which is beneficial for subsequent models to learn more refined features. IAFNet has two prediction branches. A segmentation head is used to get the final segmentation result and calculate the loss. A steering angle prediction branch, which predicts the current steering angle of the vehicle in the picture, and calculates the loss with the ground truth.

B. Steering Angle Information and Image Feature Fusion

1) *Image Feature Fetching*: During passing the image through the backbone, we get multiple image feature maps. Feature maps of different levels and scales often contain different image features and semantic information, which is valuable for dense prediction tasks such as segmentation.

2) *Steering Angle Information Fetching*: In order to better fuse image feature maps with steering angle feature maps by LIA-CAM, we need to expand the steering angle to some extent, and the expanded dimension needs to match the dimension of the image feature map. If we expand the steering angle to the same dimension as the image feature map, the number of parameters to be learned is huge, which will make the model very large and slow down the training speed.

$$A = FC_1(a) \otimes FC_2(a) \otimes FC_3(a), \quad (1)$$

where a , A , FC , $\otimes$ denotes the steering angle, angle feature maps, fully connected layer, and tensor outer product, respectively. Specifically, we expand the steering angle by passing fully connected layers three times. If the dimension of the

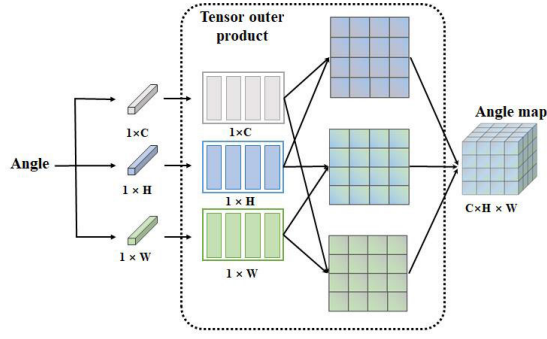


Fig. 3. Process diagram of tensor outer product. We first expand the angle in three $C \times H \times W$ dimensions, then leverage the multiplication of the three expanded tensors separately. Next, we perform a spatial tensor fusion of the obtained two-dimensional tensor maps.

image feature map is $C \times H \times W$, we expand the steering angle by three fully connected layers into three feature vectors with dimensions of $1 \times C$, $1 \times H$, and $1 \times W$, and then do the tensor outer product of the three feature vectors to obtain a $C \times H \times W$ three-dimensional feature tensor of the steering angle. The process diagram is shown in Fig. 3. In other words, we use a very small computational cost to greatly reduce the number of parameters to be learned and improve the training speed and the degree of fusion of images and steering angles.

3) *Lightweight Image-Angle Cross-Attention Module*: We perform LIA-CAM on the image feature maps and the steering angle feature maps with the same shape, using the steering angle feature maps as key sequence and image feature maps as query sequence and value sequence. It leverages the cross-attention mechanism to establish the soft associations between the steering angle and image features, enabling the neural network to adaptively determine the location and content of the information extracted from the images, resulting in an effective fusion strategy. The specific flow of the LIA-CAM is illustrated in Fig. 4.

Firstly, we perform sequentialization for steering angle feature maps A and image feature maps I . The image feature size is $H \times W \times C$. On the one hand, convolution with convolution kernel size 1×1 and reshape operation is performed on I to obtain Q sequence of size $HW \times d$. On the other hand, I is downsampled, convolved with a convolutional kernel size 1×1 , and reshaped to obtain a V sequence of size $hw \times d$. The steering angle feature size is $hw \times d$ as K sequence. The inputs to the LIA-CAM are K , Q , V , which are computed as:

$$K = AW_K, \quad Q = IW_Q, \quad V = IW_V, \quad (2)$$

where W_K is learnable parameters of the steering angle linear layer, W_Q , W_V is learnable parameters of the linear layer of the image feature map.

In Eq. 2, as we can see, we set the steering angle feature maps as K , for this setting corresponds to our motivation of needing the steering angle information for segmenting lanes in images. We perform a multiplication operation on the Q and K^T to obtain a new feature map Z . Z is calculated as $Z = QK^T$, $Z \in \mathbb{R}^{HW \times hw}$. The feature map Z contains the steering angle information and semantic information of the image, and the steering angle feature information can guide

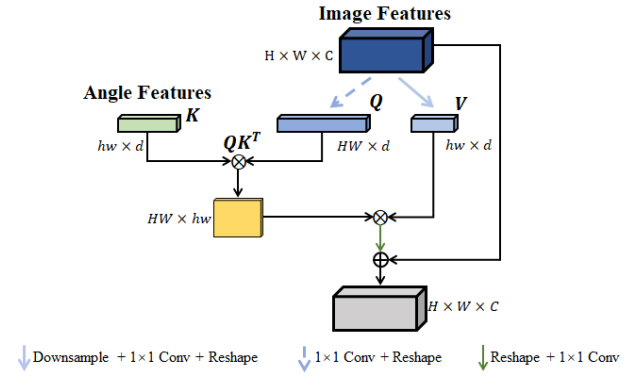


Fig. 4. Framework of LIA-CAM. The module has image features as Q , V and the angle features as K . Firstly, K , Q does matrix multiplication to get QK^T feature map. Secondly, the obtained feature map does matrix multiplication with V . Finally, the result is summed with image features to get the final output.

the network to learn the semantic information of the lane lines in the image more comprehensively.

Secondly, in order to learn the steering angle information at a deeper level, we need to combine Z with the value sequence V to obtain a deep angle information image feature map Z' , $Z' \in \mathbb{R}^{H \times W \times C}$. The combination process first multiplies Z and V , and then performs a reshape operation and a 1×1 convolution operation on the result obtained in the previous step. This process can fully combine the steering angle information with the image feature information and restore the learned feature map to the original size. Z' satisfies the Eq. 3:

$$Z' = \text{Conv}[\text{Reshape}(Z \times V)], \quad (3)$$

Finally, there is a residual connection between the deep angle information image feature map Z' and the original input image feature map I to obtain the LIA-CAM output feature map O , which is formulated as Eq. 4:

$$O = QK^T \times V + I, \quad (4)$$

where O size is $H \times W \times C$. The feature map O combines the steering angle information more comprehensively and deeply, which can assist the network to learn the image feature information more stably and effectively.

C. Loss Function

We propose a combined loss function to supervise the two sub-tasks of segmentation and steering angle prediction, respectively.

$$L = \frac{1}{2\sigma_1^2}(L_{ce} + L_{dice}) + \frac{1}{2\sigma_1^2}L_{mse} + \log\sigma_1 + \log\sigma_2, \quad (5)$$

The lane line segmentation has the property of severe class imbalance, where the foreground class (i.e., lane line class) accounts for only 2 percent of all pixels. Therefore, using the cross-entropy loss function [54] alone will get not good results. Here we use a combination of cross-entropy loss function and Dice loss to calculate the segmentation loss. In addition to this, we use the mean square error loss to calculate the loss

TABLE I

COMPARISON WITH COMMON DATASETS IN TERMS OF ANNOTATED DATA AND SCENES, WHERE FEW, MEDIUM, AND LARGE INDICATE THE PROPORTION OF THAT SCENE IN THE ALL DATASET

Dataset	Annotations		Scenes				
	Angle	Lane	Strong Light	Dim Light	Slight Turns	Rapid Turns	Straight Lane
Caltech Lanes [55]		✓	Few	Few	Few	Few	Large
TuSimple [56]		✓	Medium	Few	Large	Few	Large
CULane [57]		✓	Large	Medium	Large	Medium	Large
OpenMPD [58]		✓	Large	Few	Large	Medium	Large
BDD100K [59]		✓	Large	Large	Large	Medium	Large
IA-Udacity (ours)	✓	✓	Large	Large	Large	Large	Large

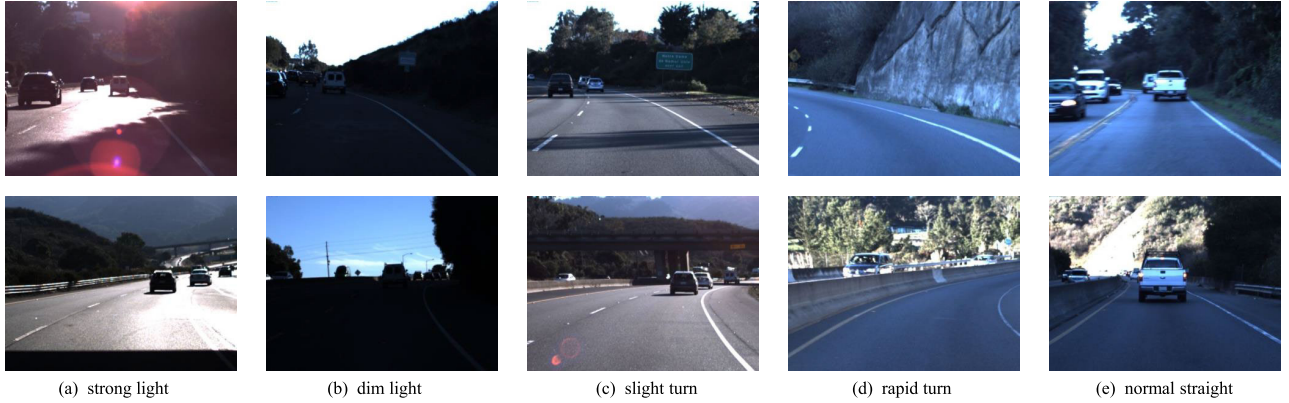


Fig. 5. The ImageAngle-Udacity dataset includes examples with: (a) strong light and (b) dim light, which affect lane detection due to lighting effects, as well as (c) slight turns, (d) rapid turns, and (e) normal straight driving, which affect lane detection due to angle effects.

of vehicle steering angle prediction, and our motivation is to motivate the network to keep the correct vehicle steering angle information from the beginning to the end by using this loss function. To deal with the problem of large differences in the scales of the two loss values of segmentation loss and regression loss, we use a method to automatically learn the loss weights for different tasks. They derived a principled loss function that can learn relative weighting automatically from the data and is robust to the weight initialization.

IV. EXPERIMENTS

Due to the common lane detection datasets not including steering angle information, we have constructed a new dataset based on the Udacity dataset that includes both images and steering angle information, named IA-Udacity. In the section IV-A, we introduce the proposed IA-Udacity. In the section IV-B, we describe the implementation details of the experiments. In the section IV-C, we select multiple previous lane line detection methods for comparison experiments. In the section IV-D, we perform ablation experiments for our proposed Angle-Image Fusion strategy (the number and position of fusion) and multi-task losses. Finally, we discuss the limitations of the steering angle and its potential impact on future fields in section IV-F.

A. Datasets

Table I shows the data annotations and scene coverage of existing public datasets, such as Caltech Lanes [55], TuSimple [56], CULane [57], OpenMPD [58], and BDD100K [59]. It is evident that these datasets primarily focus on lane

marking annotations and do not include crucial steering angle information, which can be instrumental for more advanced lane detection and driving assistance systems. Moreover, these datasets exhibit limited coverage in challenging driving conditions such as strong light, dim light, and rapid turns. In stark contrast to existing methods, our approach aims to leverage steering angle information from vehicles to enhance lane detection, thereby providing significant prior knowledge that improves the robustness and accuracy of lane detection systems. Therefore, we have constructed a dataset based on the Udacity dataset [60], named ImageAngle-Udacity (IA-Udacity),¹ which includes images, steering angles, and lane markings. Our dataset is specifically designed to encompass a wide range of challenging scenarios, including road images captured under various lighting conditions (strong light, dim light, and normal light) and featuring a range of turning angles (slight turns, sharp turns, and straight lane), as illustrated in Fig. 5. By incorporating steering angle data and ensuring extensive coverage of difficult driving scenarios, our dataset offers a more comprehensive tool for the development and evaluation of advanced lane detection algorithms. This not only fills the gaps left by previous datasets but also offers new insights for research in the fields of autonomous driving and ADAS.

The Udacity dataset [60] is composed of video frames taken from urban roads, which provides a total number of 404,916 video frames for training and 5,614 video frames for testing. Specifically, we select the CH20_02 part in the Udacity dataset. It contains five driving scenarios with a total

¹<https://iee-dataport.org/documents/imageangle-udacity-ia-udacity>

TABLE II
COMPARISON WITH PREVIOUS STATE-OF-THE-ART
METHODS ON IA-UDACITY

Methods	IoU	F1-score	Precision	Recall
LaneNet [12]	73.77%	84.69%	86.88%	82.60%
SCNN [13]	74.92%	84.88%	86.92%	82.93%
RESA [14]	75.20%	85.30%	87.10%	84.40%
LDNet [61]	76.12%	85.87%	88.04%	83.80%
ConvNeXt [62]	73.75%	84.36%	86.23%	82.49%
FASeg [63]	72.65%	84.16%	85.97%	82.35%
SwinT [45]	71.13%	83.53%	85.68%	81.38%
Ours	77.33%	86.59%	87.27%	85.91%
Ours+angle	78.41%	87.99%	89.01%	86.95%

of 1,694 seconds of video data. We select one image and its corresponding steering angle per 10 frames to get a total of 6,308 images (640×480). Each of these images contains a normalized turning angle within the range of -1 to 1 , with negative numbers representing right turns and positive numbers representing left turns. After fine labeling, the masks of lane lines are obtained. Then we randomly divide the training set and testing set in a ratio of 7:3, with a total of 4,415 images in the training set and 1,893 images in the testing set.

B. Implementation Details

In the process of network training and testing, the input is a three-channel image of size 640×480 . For the steering angle, we first transform the value type into a one-dimensional tensor and then extend its dimensionality to be the same as the feature map dimension, which helps the network to learn effectively. The network training epoch is set to 150 and the batch sizes and max training iterations are 4 and 80k. We use the Stochastic Gradient Descent (SGD) optimizer algorithm with an initial learning rate of $1e-3$, momentum of 0.9, and weight decay of $5e-4$ to train IAFNet. Besides, a poly learning rate policy is adopted. We build our model with PyTorch toolkit and all experiments are conducted on one RTX 3090 GPU card with 24 GB memory.

C. Quantitative Results

In order to verify the effectiveness of the proposed method, we compare our model with a variety of lane line detection methods, including models such as LaneNet [12], SCNN [13], RESA [14], LDNet [61], ConvNeXt [62], FASeg [63] and SwinT [45]. The comparison metrics include IoU, F1-score, Precision, and Recall, towards comprehensively reflecting the performance of the models. The comparison results are shown in Table II. It can be seen that our model achieves the best results, IAFNet's IoU is 78.41% and F1-score is 87.99%, outperforming the second place model in IoU and F1-score by 2.29% and 2.12%, respectively. Comparing our method without the steering angle feature information with the steering angle feature information, IoU and F1-score are improved by 1.07% and 1.4%, respectively. It demonstrates that by utilizing the implied relationship between the steering angle and image features, the steering angle can guide the network to

TABLE III
COMPARISON OF PERFORMANCE BETWEEN THE LEADING LANE DETECTION METHOD WITH AND WITHOUT THE STEERING ANGLE

Method	IoU	F1-score	Precision	Recall
GANet [64]	75.45	85.26	86.81	83.76
GANet-Angle	76.28	85.91	87.42	84.93
CLRNet [65]	75.14	84.91	86.73	83.17
CLRNet-Angle	76.09	85.83	87.46	84.26
BézierLaneNet [66]	76.12	85.47	86.27	84.68
BézierLaneNet-Angle	77.33	86.60	87.70	85.54
PGA-Net [67]	76.79	86.42	87.38	85.49
PGA-Net-Angle	77.61	87.57	88.57	86.60

learn image features more efficiently and improve the network performance.

To verify the benefits of steering angle information for lane detection algorithms, we have incorporated the proposed angle branch into GANet [64], CLRNet [65], BézierLaneNet [66], and PGA-Net [67], and utilized LIA-CAM on the backbone to adjust the lane feature map representation, as shown in Table III. It can be seen that incorporating steering angle information results in an approximate 1% performance improvement in terms of IoU for each method. The experimental results demonstrate that the steering angle can provide rich prior knowledge for lane detection algorithms, significantly improving the accuracy of lane detection. Furthermore, acquiring angle information is nearly zero-cost for vehicles, which holds tremendous potential for applications.

In order to verify the effectiveness of the proposed LIA-CAM, we compare with current state-of-the-art attention, including Squeeze-and-Excitation Networks (SE) [68], Efficient Channel Attention (ECA) [69], External Attention (EA) [70], Pyramid Squeeze Attention (PSA) [71], Convolutional Block Attention Module (CBAM) [72], and Feature Fusion Module (FFM) [73]. The experimental results of the comparison are shown in Table IV. It can be seen from the table that LIA-CAM outperforms the other attention modules, achieving 78.41% and 87.99% on IoU and F1-score, respectively. Compared to the best attention mechanism CBAM, LIA-CAM IoU and F1-score are 4.32% and 2.50% higher than CBAM. It demonstrates that the steering angle can provide prior knowledge for lane line detection. The LIA-CAM that fuses the steering angle features with image features can improve network performance. Compared with the latest FFM [73], our proposed LIA-CAM outperforms it by 2.87% in terms of IoU. This is because FFM assumes that both modalities are equally important by default. In stage 1 of FFM, it performs self-attention operations separately within the two modalities. However, the angle information itself, being a low-dimensional tensor, does not possess any positional information and only provides certain prior knowledge of the image features.

D. Qualitative Results

As shown in Fig. 6, our network is more accurate in detecting the edge part of the lane lines compared to other networks under harsh environments. Under poor lighting conditions, it can be observed that our network is more accurate for lane line detection. Especially for the lane line segmentation at the

TABLE IV
COMPARISON OF PERFORMANCE WITH THE PREVIOUS STATE-OF-THE-ART ATTENTION MECHANISM METHODS

Attention	IoU	F1-score	Precision	Recall	#Parameters (K)
SE [68]	71.09%	82.57%	83.26%	81.88%	2.048
ECA [69]	72.31%	83.56%	84.49%	82.63%	0.003
EA [70]	72.03%	83.41%	84.37%	82.45%	1.638
PSA [71]	73.28%	84.47%	85.68%	83.26%	3.725
CBAM [72]	74.09%	85.49%	86.74%	84.24%	4.467
FFM [73]	75.54%	86.73%	88.05%	85.36%	15.048
LIA-CAM (ours)	78.41%	87.99%	89.01%	86.95%	1.722

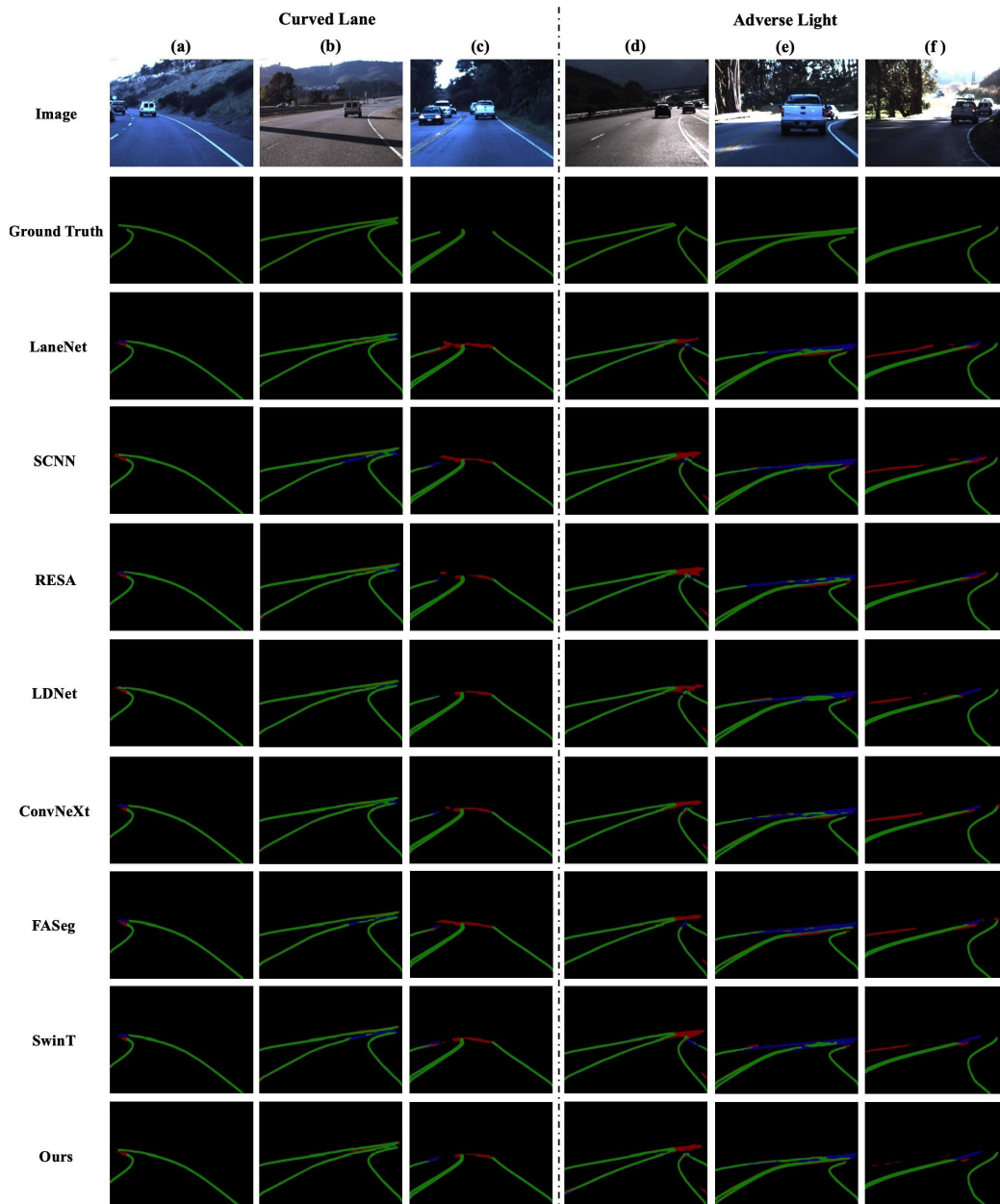


Fig. 6. Qualitative results on IA-Udacity dataset. (a)-(c) is the images of the curved lane line. (d)-(e) is the images of the adverse light. Green masks are true positives, red masks are false positives, and blue masks are false negatives. Both over-segmentation and under-segmentation existed in the predicted segmentation maps of each model.

far end of the road, our model can accurately segment the edge of the lane line. Meanwhile, in the case of vehicle occlusion, our model can accurately detect the lane lines in the occluded part. When the road is under adverse light, we find it difficult

to distinguish the location of the lane lines, but the network can learn the location of the lane lines. We also observe that our model learns the detailed features of the lane lines in the image. These demonstrate that networks fusing steering angle

TABLE V

ABLATION STUDIES OF THE IMAGE-ANGLE FUSION STRATEGY. S2, S3 AND S4 REPRESENT THE FUSION OF STEERING ANGLE INFORMATION AND IMAGE INFORMATION IN THE 2ND, 3RD AND 4TH STAGES OF THE DECODER, RESPECTIVELY

S2	S3	S4	IoU	F1-score	Precision	Recall
			77.33%	86.59%	87.27%	85.91%
✓			77.39%	86.00%	87.30%	84.70%
	✓		77.41%	86.62%	87.38%	85.86%
		✓	77.38%	86.61%	87.31%	85.91%
✓	✓		77.80%	87.15%	88.41%	85.89%
✓		✓	77.72%	87.04%	88.36%	85.72%
	✓	✓	77.83%	87.26%	88.49%	86.05%
✓	✓	✓	78.41%	87.99%	89.01%	86.95%

features and image features can learn the implicit relationship between steering angle features and image features to improve the accuracy of detecting lane lines.

E. Ablation Study

We conduct ablation experiments on the IA-Udacity dataset generally in three parts: the image-angle fusion strategy (including the number and positions of fusion), the multi-task loss, the steering angle embedding and LIA-CAM to demonstrate their effectiveness.

The results of the ablation experiment for the fusion strategy are shown in Table V. We perform the experiments from two perspectives, the fusion number and the fusion position respectively, and it can be seen from Table V that the best segmentation performance can be obtained by fusing the steering angle features with the image features in stages 2, 3 and 4. In the case of single-stage fusion, the network performance does not gain a significant improvement compared to the network structure without the steering angle feature. In the case of two-stage fusion, we can find the best fusion effect in stages 3 and 4. Compared with the best fusion effect in a single stage, the IoU and F1-score are improved by 0.42% and 0.64%, respectively. The IoU and F1-score are 77.83% and 87.26% for the convergence strategy in stages 3 and 4, respectively. The fusion strategy of stages 3 and 4 is superior compared to the fusion strategy of stages 2 and 3 and the fusion strategy of stages 2 and 4. It illustrates that combining steering angle features with low-level semantic features of the images is more accurate for the network to learn lane line locations. We also observe that the three-stage fusion strategy results in the best segmentation performance of the network. The IoU and F1-score are 78.41% and 87.99%, respectively, mainly attributed to IAFNet. The three-stage fusion strategy allows us to have a comprehensive fusion learning of steering angle features and image features. The feature learning of lane lines in images can be more accurate with the assistance of steering angle features.

To illustrate the importance of the multi-task dynamically adjusted weight loss function, we conducted ablation experiments of segmentation loss, regression loss, and multi-task dynamically adjusted weight loss. We also did experiments incorporating steering angle information into our network and training steering angle information features using a regression

TABLE VI

ABLATION STUDIES OF THE MULTI-TASK LOSS FUNCTIONS. SEGLOSS IS THE SEGMENTATION LOSS. REGLOSS IS THE REGRESSION LOSS. REWEIGHT IS THE DYNAMICALLY ADJUSTED WEIGHT LOSS. ANGLE IS THE STEERING ANGLE FEATURES

SegLoss	RegLoss	Reweight	Angle	IoU	F1-score	Precision	Recall
✓				77.33%	86.59%	87.27%	85.91%
✓	✓			77.57%	86.89%	87.32%	86.46%
✓	✓	✓		77.99%	87.23%	87.73%	86.73%
✓	✓	✓	✓	78.41%	87.99%	89.01%	86.95%

TABLE VII

ABLATION STUDIES OF THE DIFFERENT SCENES. SCENES CURVED LANE LINE SCENES AND BAD LIGHTING SCENES OF TWO KINDS. THE EXPERIMENTS ARE PERFORMED ON TWO MODELS WITH AND WITHOUT STEERING ANGLE INFORMATION

Scenes	Angle	IoU	F1-score	Precision	Recall
curved lane	ours	75.78%	84.54%	85.17%	83.91%
	ours+angle	76.59%	85.04%	85.99%	84.09%
adverse light	ours	74.32%	83.68%	83.96%	83.40%
	ours+angle	75.42%	84.08%	84.68%	83.51%

loss function. The experimental results are shown in Table VI. The IoU and F1-score of our model reach 77.33% and 86.59%, respectively, if the segmentation loss is used alone. If the regression loss is added, the IoU and F1-score of our model improve by 0.24% and 0.40%, respectively. In our model, the IoU and F1-score are 77.99% and 87.23%, respectively, when using both segmentation loss, regression loss, and multi-task dynamically adjusted weight loss. When we incorporate the steering angle information into our network and train the steering angle information features using regression loss, our model IoU and F1-score are 78.41% and 87.99%, respectively. Compared to the loss method using a combination of all three, the IoU and F1-score of the model are improved by 0.42% and 0.76%, respectively. It can be seen that by introducing the regression loss for the steering angle, the segmentation performance of the model can be improved noticeably. We believe this is because the presence of the steering angle regression loss can induce the network model to always retain the correct steering angle value, which in turn ensures the performance improvement resulting from the fusion of steering angle features with image features.

To verify the effectiveness of the fusion of angle features and image features in different scenes, we present the results of their ablation experiments in the Table VII. In the case of curved lane lines, the model with steering angle features fused with image features is superior to the model without corners. The IoU and F1-score of the two combined networks are 76.59% and 85.04%, respectively. Compared to the model without steering angle features, the IoU and F1-score of the fused model with angle features and image features are higher by 0.81% and 1.1%. In the case of adverse light, the IoU and F1-score of the model with steering angle features fused with image features 75.42% and 84.08%, respectively. Compared to the model without steering angle features, the IoU and F1-score of the fused model with angle features and image features are higher by 1.10% and 0.40%. These indicate that the network with the steering angle information can learn the

TABLE VIII

ABLATION STUDIES OF THE STEERING ANGLE EMBEDDING AND LIA-CAM. TE IS THE ORDINARY TENSOR EXPANSION. CA IS THE ORDINARY CROSS-ATTENTION. TPF IS THE TENSOR OUTER PRODUCT FUSION

TE	CA	TPF	LIA-CAM	Params (M)	FLOPs (G)	FPS
✓	✓			64.21	240.45	19
✓			✓	60.57	237.21	24
	✓	✓		59.43	235.10	25
		✓	✓	56.39	233.10	29

implicit relationship between the steering angle features and image features.

To verify the effectiveness of the combination of the steering angle and LIA-CAM, we present the results of their ablation experiments in the Table VIII. From Table VIII, we can observe that the model complexity of combining the steering angle and LIA-CAM into one, i.e., floating point operations (FLOPs) is 233.10, and frames per second (FPS) is 29. When we add the steering angle features and LIA-CAM to the model, the parameters and complexity of the model are reduced instead, and the model is computationally faster. This proves that IAFNet can not only reduce the number of parameters and the FLOPs but also enhance the FPS and the performance of lane detection.

F. Discussion

To the best of our knowledge, our paper is the first to propose the use of steering angle to assist lane detection algorithms based on RGB cameras. Therefore, we wish to further discuss the limitation of steering angle and its potential impact on the field of lane detection and autonomous driving.

1) *Our Limitation:* Based on our observation of the videos from the Udacity dataset [60], the driver's maneuvers are relatively stable. Although the steering angle information obtained by drivers with limited steering control capabilities may not be accurate, the general direction is certain. The steering angle information merely assists lane detection in the form of weak supervision, without causing significant performance degradation in this situation. However, if the driver engages in distracted driving, frequently changes lanes at will, or exhibits malicious and irregular lane departure behaviors, such prior knowledge could have a detrimental effect on lane detection.

2) *Our Potentiality:* Our paper presents the idea of incorporating steering angle information into lane detection, which has a significant impact on the field of lane detection. This idea has been validated through extensive experimentation, yet this is just the beginning. What we wish to convey is that vehicle information is of crucial importance to the entire field of autonomous driving, encompassing all states of the vehicle, including but not limited to steering angle, vehicle speed, electric current, and torque. However, the majority of existing autonomous driving methods focus on external sensors such as radar, LiDAR, ultrasonic, and cameras, neglecting the most reliable and cost-effective vehicle information. We firmly believe that among the various existing methods, the integration of vehicle information will undoubtedly produce a variety of remarkable effects. In the future, we plan to design

a large-scale autonomous driving perception model by integrating vehicle information, such as steering angle and speed, along with sensor data from cameras, radar, LiDAR, and other sensors. This approach aims to adequately address sensor failure scenarios, enhancing the overall safety and reliability of the autonomous driving system.

V. CONCLUSION

In this paper, a novel network with multimodal fusion that combines steering angle features and image features was proposed. The proposed method solves the problem of curved lane line detection in complex environments. Firstly, we introduce steering angle information and fuse the steering angle information with image feature information to assist lane detection. Secondly, IAFNet is an encoder-decoder network structure that exploits the implied relationship between multi-level steering angle features and low-level or high-level image features by LIA-CAM, fusing multi-level steering angle features with multi-level image features multiple times. Finally, we have released a new dataset: IA-Udacity, which contains some of the images in Udacity. IAFNet achieved state-of-the-art performance on this dataset, surpassing other existing methods.

REFERENCES

- [1] A. Lindgren and F. Chen, "State of the art analysis: An overview of advanced driver assistance systems (ADAS) and possible human factors issues," *Hum. Factors Econ. Aspects Saf.*, pp. 38–50, Apr. 2006.
- [2] Z. Bairi, O. Ben-Ahmed, A. Amamra, A. Bradaï, and K. Beghdad Bey, "PSCS-net: Perception optimized image reconstruction network for autonomous driving systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 1564–1579, Feb. 2023.
- [3] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 9961–9980, Aug. 2021.
- [4] Y. Gong et al., "TCLaneNet: Task-conditioned lane detection network driven by vibration information," *IEEE Trans. Intell. Veh.*, early access, Jan. 22, 2024, doi: [10.1109/TIV.2024.3356813](https://doi.org/10.1109/TIV.2024.3356813).
- [5] H. Ning, R. Yin, A. Ullah, and F. Shi, "A survey on hybrid human-artificial intelligence for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6011–6026, Jul. 2022.
- [6] Y. Li, X. Tao, X. Zhang, J. Liu, and J. Xu, "Privacy-preserved federated learning for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8423–8434, Jul. 2022.
- [7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
- [9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
- [11] Y. Gong, L. Wang, and L. Xu, "A feature aggregation network for multispectral pedestrian detection," *Appl. Intell.*, vol. 53, no. 19, pp. 22117–22131, 2023.
- [12] D. Neven, B. D. Brabandere, S. Georgoulis, M. Proesmans, and L. V. Gool, "Towards end-to-end lane detection: An instance segmentation approach," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 286–291.
- [13] A. Parashar et al., "SCNN: An accelerator for compressed-sparse convolutional neural networks," in *Proc. ACM/IEEE 44th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2017, pp. 27–40.
- [14] T. Zheng et al., "RESA: Recurrent feature-shift aggregator for lane detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3547–3554.

- [15] X. Li, J. Li, X. Hu, and J. Yang, "Line-CNN: End-to-end traffic line detection with line proposal unit," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 248–258, Jan. 2020.
- [16] L. Tabelini, R. Berriel, T. M. Paixao, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "Keep your eyes on the lane: Real-time attention-guided lane detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 294–302.
- [17] Z. Tan, J. Feng, and J. Zhou, "SGNet: Structure-aware graph-based network for airway semantic segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2021, pp. 153–163.
- [18] J. Zhang, T. Deng, F. Yan, and W. Liu, "Lane detection model based on spatio-temporal network with double convolutional gated recurrent units," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6666–6678, Jul. 2022.
- [19] Z. Yao, X. Wu, P. Wang, and C. Ding, "DevNet: Deviation aware network for lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17584–17593, Oct. 2022.
- [20] G. Pang, B. Zhang, Z. Teng, N. Ma, and J. Fan, "Fast-HBNet: Hybrid branch network for fast lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15673–15683, Sep. 2022.
- [21] Y. Gong et al., "SIFDriveNet: Speed and image fusion for driving behavior classification network," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 1, pp. 1244–1259, Feb. 2024.
- [22] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, *arXiv:1707.07250*.
- [23] G. Francopoulo et al., "Lexical markup framework (LMF)," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, May 2006, pp. 233–236.
- [24] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558.
- [25] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 14200–14213.
- [26] X. Zhang et al., "Multi-modal fusion technology based on vehicle information: A survey," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 6, pp. 3605–3619, Jun. 2023.
- [27] X. Bai et al., "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1090–1099.
- [28] Y. Cao, H. Zhang, Y. Li, C. Ren, and C. Lang, "CMAN: Leaning global structure correlation for monocular 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24727–24737, Dec. 2022.
- [29] D. Zhou, X. Song, J. Fang, Y. Dai, H. Li, and L. Zhang, "Context-aware 3D object detection from a single image in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18568–18580, Oct. 2022.
- [30] Q. He et al., "Stereo RGB and deeper LiDAR-based network for 3D object detection in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 152–162, Jan. 2023.
- [31] Y. Zhang, Z. Lu, X. Zhang, J.-H. Xue, and Q. Liao, "Deep learning in lane marking detection: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 5976–5992, Jul. 2022.
- [32] Y. Gong et al., "SkipcrossNets: Adaptive skip-cross fusion for road detection," 2023, *arXiv:2308.12863*.
- [33] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7482–7491.
- [34] Y. Gong, "ImageAngle-Udacity (IA-Udacity)," Tsinghua Univ., Beijing, China, IEEE Dataport, Mar. 2023, doi: [10.21227/1evd-ew07](https://doi.org/10.21227/1evd-ew07).
- [35] L. Liu, X. Chen, S. Zhu, and P. Tan, "CondLaneNet: A top-to-down lane detection framework based on conditional convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3773–3782.
- [36] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection CNNs by self attention distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1013–1021.
- [37] Z. Qin, H. Wang, and X. 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*. Springer, 2020, pp. 276–291.
- [38] Z. Qu, H. Jin, Y. Zhou, Z. Yang, and W. Zhang, "Focus on local: Detecting lane marker from bottom up via key point," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14122–14130.
- [39] M. Morley, R. Atkinson, D. Savić, and G. Walters, "GAnet: Genetic algorithm platform for pipe network optimisation," *Adv. Eng. Softw.*, vol. 32, no. 6, pp. 467–475, 2001.
- [40] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. Pu Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, *arXiv:1806.00064*.
- [41] M. Lipinski, T. Wlostowski, J. Serrano, and P. Alvarez, "White rabbit: A PTP application for robust sub-nanosecond synchronization," in *Proc. IEEE Int. Symp. Precis. Clock Synchronization Meas., Control Commun.*, Sep. 2011, pp. 25–30.
- [42] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.
- [43] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [44] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [45] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [46] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. Glasgow, U.K.: Springer*, Aug. 2020, pp. 213–229.
- [47] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
- [48] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.
- [49] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.
- [50] L. Bai, L. Wang, T. Chen, Y. Zhao, and H. Ren, "Transformer-based disease identification for small-scale imbalanced capsule endoscopy dataset," *Electronics*, vol. 11, no. 17, p. 2747, Aug. 2022. [Online]. Available: <https://www.mdpi.com/2079-9292/11/17/2747>
- [51] Z. Zhang et al., "Neural machine translation with universal visual representation," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–14.
- [52] Z. Zhang et al., "MAFF-net: Filter false positive for 3D vehicle detection with multi-modal adaptive feature fusion," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 369–376.
- [53] X. Dai, X. Kong, and T. Guo, "EPNet: Learning to exit with flexible multi-branch network," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 235–244.
- [54] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.
- [55] M. Aly, "Real time detection of lane markers in urban streets," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2008, pp. 7–12.
- [56] TuSimple. (2017). *TuSimple Benchmark*. Accessed: Jan. 2023. [Online]. Available: <https://github.com/TuSimple/tusimple-benchmark>
- [57] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.
- [58] X. Zhang et al., "OpenMPD: An open multimodal perception dataset for autonomous driving," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2437–2447, Mar. 2022.
- [59] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2636–2645.
- [60] E. Gonzalez, M. Virgo, J. Jensen, and C. Kirksey. (2018). *Annotated Driving Dataset*. [Online]. Available: <https://github.com/udacity/self-driving-car/tree/master>
- [61] F. Munir, S. Azam, M. Jeon, B.-G. Lee, and W. Pedrycz, "LDNet: End-to-end lane marking detection approach using a dynamic vision sensor," 2020, *arXiv:2009.08020*.
- [62] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11976–11986.
- [63] H. He et al., "Dynamic focus-aware positional queries for semantic segmentation," 2022, *arXiv:2204.01244*.
- [64] J. Wang et al., "A keypoint-based global association network for lane detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1392–1401.

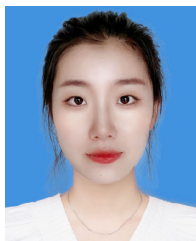
- [65] T. Zheng et al., “CLRNet: Cross layer refinement network for lane detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 898–907.
- [66] Z. Feng, S. Guo, X. Tan, K. Xu, M. Wang, and L. Ma, “Rethinking efficient lane detection via curve modeling,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17062–17070.
- [67] Q. Li et al., “PGA-net: Polynomial global attention network with mean curvature loss for lane detection,” *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 417–429, Jan. 2024.
- [68] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [69] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-net: Efficient channel attention for deep convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
- [70] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, “Beyond self-attention: External attention using two linear layers for visual tasks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5436–5447, May 2023.
- [71] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, “EPSANet: An efficient pyramid squeeze attention block on convolutional neural network,” 2021, *arXiv:2105.14447*.
- [72] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
- [73] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, “CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers,” *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.



Yan Gong received the M.S. degree in computer science and technology from Northeastern University, Shenyang, China, in 2023. He was a joint training Master’s Student with the School of Vehicle and Mobility, Tsinghua University, Beijing, China, where he was also a Research Assistant. He is currently a Researcher with the Autonomous Driving Division Department of X Division, JD Logistics. He has authored over ten academic papers. His research interests include pattern recognition, multi-modal fusion, and deep learning.



Xinyu Zhang (Member, IEEE) received the B.E. degree from the School of Vehicle and Mobility, Tsinghua University, Beijing, China, in 2001, and the M.S. degree from Tsinghua University in 2004. He was a Visiting Scholar with the University of Cambridge. He is currently an Associate Professor with the School of Vehicle and Mobility, and the Head of the Mengshi Intelligent Vehicle Team, Tsinghua University. He is the author of more than 50 articles. His research interests include intelligent driving and multimodal information fusion.



Jianli Lu received the degree in computer technology from China University of Mining and Technology (Beijing), China. She has been a Research Assistant with the School of Vehicle and Mobility, Tsinghua University, China. Her research interests include computer vision, multi-modal fusion, and deep learning.



Xinmin Jiang is currently pursuing the bachelor’s degree in computer science and technology with Northeastern University, China. Her research interests include autonomous driving, multi-modal fusion, and deep learning.

Zichen Wang, photograph and biography not available at the time of publication.

Hao Liu, photograph and biography not available at the time of publication.



Zhiwei Li is a Tutor of master’s students with Beijing University of Chemical Technology. In 2020, he studied as a Post-Doctoral Fellow with Academician Jun Li with Tsinghua University. His main research interests include computer vision, intelligent perception and autonomous driving, and robot system architecture.



Li Wang received the Ph.D. degree in mechatronic engineering from the State Key Laboratory of Robotics and System, Harbin Institute of Technology, in 2020. He was a Visiting Scholar with Nanyang Technology of University, for two years. He was a Post-Doctoral Fellow with the State Key Laboratory of Automotive Safety and Energy and the School of Vehicle and Mobility, Tsinghua University. Currently, he is an Assistant Professor with the School of Mechanical Engineering, Beijing Institute of Technology. He is the author of more than 30 SCI/EI articles. His research interests include autonomous driving perception, 3-D robot vision, and multi-modal fusion.

Qingshan Yang, photograph and biography not available at the time of publication.



Xingang Wu was a Software Engineer with the Information Technology Center, Tsinghua University, from 2007 to 2018. In 2015, he joined the Chinese Society for Artificial Intelligence, as a member. Currently, he is a Software Engineer with the School of Vehicles and Mobility, Tsinghua University. His current research interests include software engineering, intelligent vehicle safety theory, and Internet of Things security theory.