Automotive Innovation (2024) 7:545–558
https://doi.org/10.1007/s42154-024-00296-x

MFE-SSNet: Multi-Modal Fusion-Based End-to-End Steering Angle and Vehicle Speed Prediction Network

Yi Huang[1] · Wenzhuo Liu[2] · Yaoyu Li[1] · Lei Yang[1] · Hanqi Jiang[3] · Zhiwei Li[4] · Jun Li[1]

Received: 6 September 2023 / Accepted: 26 February 2024 / Published online: 24 October 2024

Abstract
In the field of autonomous vehicles, accurately predicting steering angle and speed is a pivotal task. This task affects the accuracy of the final decision of the autonomous vehicle and is the basis for ensuring the safe and efficient operation of the autonomous vehicle. Previous studies have often relied on data from only one or two modalities to make predictions for steering angle and vehicle speed, which were often inadequate. In this paper, the authors propose a Multi-Modal Fusion-Based End-to-End Steering Angle and Vehicle Speed Prediction Network (MFE-SSNet). The network innovatively extends the one-stream and two-stream structure to a three-stream structure and cleverly extracts features of images, steering angles, and vehicle speeds using HRNet and LSTM layers. In addition, in order to fully fuse the feature information of different modal data, this paper also proposes a local attention-based feature fusion module. This module improves the fusion of different modal feature vectors by capturing the interdependencies in the local channels. Experimental results demonstrate that MFE-SSNet outperforms the current state-of-the-art model on the publicly available Udacity dataset.

Keywords Autonomous vehicles · Deep learning · Multimodal fusion · Intelligent transportation

1 Introduction

The realm of the automotive industry is perpetually evolving, and a key factor propelling this progression is the development and implementation of predictive technologies. The prediction of optimal driving speed and steering angle has emerged as an especially pivotal technology in this regard. These predictions, utilizing vast quantities of vehicle sensor data and advanced predictive algorithms, enable the precise forecasting of ideal driving conditions, which results in substantial enhancements in driving safety, autonomous driving technology, and energy efficiency [1]. The ramifications of these advantages span multiple dimensions, encompassing the improvement of driving experiences, diminution of traffic accident risks, augmentation of fuel efficiency, and the advocacy of environmental sustainability [2]. Consequently, predictive technology plays an indispensable role in the transition towards intelligent driving and sustainable transportation systems, solidifying its position as a cornerstone of the industry.

Despite recent advancements in autonomous driving technology, the field still faces significant challenges. While novel end-to-end autonomous driving models that leverage deep neural networks for both training and testing phases have been proposed [3, 4], these models are limited in their prediction scope, primarily forecasting only the subsequent five frames of vehicle speed or steering angle. This makes them less practical and challenging to apply in real-life scenarios where longer-term predictions are necessary. Additionally, these models do not sufficiently analyze the complex relationships between different input frames and outputs, which is a crucial aspect for achieving comprehensive prediction effectiveness. Another challenge is that these models tend to rely on modal fusion methods that overlook multi-dimensional modality fusion. This over-reliance on simplistic fusion methods results in sub-optimal fusion

✉ Jun Li
lijun1958@tsinghua.edu.cn

1 School of Vehicle and Mobility, Tsinghua University, Beijing, China

2 School of Electromechanical and Information Engineering, China University of Mining and Technology, Beijing, Beijing, China

3 School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

4 College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China

effectiveness, as complex inter-dependencies across multiple dimensions are not taken into account. Moreover, these methods are associated with high costs and sub-optimal predictive performance, leading to an unsatisfactory trade-off between cost and performance.

As a result, there is a pressing need for continued research and innovation in this field to address these challenges and improve the practicality and applicability of these models in real-world scenarios. Future developments should focus on improving the prediction scope and analyzing complex relationships between inputs and outputs, while also exploring more effective and efficient multi-dimensional modality fusion methods. Such advancements have the potential to enhance the robustness, adaptability, and safety of autonomous driving systems, bringing us closer to a fully autonomous future.

In response to these challenges, this study introduces the MFE-SSNet, a novel approach to predicting optimal driving behaviors in autonomous vehicles. This approach is centered around the Local Attention-based Feature Fusion Module (LA-FFM), a unique strategy that enables optimal interactions between each feature and its adjacent counterparts, thereby maintaining model complexity within feasible limits [5]. The focus on local feature interactions is critical for learning effective feature attention and avoiding dimensionality reduction, which in turn boosts performance while preserving model simplicity. Notably, the LA-FFM module displays robust performance even in sparse scenarios where direct feature interactions are minimal.

Building on the three-stream network architecture identified in previous research as beneficial for predicting vehicle speed and steering angles [6], the authors propose a tri-stream MFE-SSNet. This network efficiently extracts features from image, velocity, and steering angle flows, utilizing HRNet and LSTM layers. By integrating the LA-FFM mechanism, the MFE-SSNet enhances the accuracy and robustness of speed and steering angle predictions, especially in sparse settings. Moreover, MFE-SSNet exhibits excellent adaptability in sparse environments, outperforming other models in predicting speed and steering angles.

The experimental results provide strong evidence for the effectiveness of the MFE-SSNet. Not only does the model demonstrate robustness in sparse scenarios, but it also consistently outperforms traditional fusion strategies in terms of both accuracy and robustness. The accurate environmental predictions made by the model have profound implications for the future of autonomous driving technologies, including improvements in road safety, driving experiences, and sustainability. Therefore, this research contributes valuable insights to the autonomous driving field, opening up new avenues for research and laying a solid foundation for future studies. The overall contributions of this paper are listed as follows:

- This paper proposes a novel MFE-SSNet model, a fresh approach to predicting optimal driving data like steering angle and vehicle speed in autonomous driving. The designed tri-stream network model excels in extracting features efficiently from image, vehicle speed, and steering angle streams, resulting in the exceptional enhancement of prediction accuracy for both steering angle and vehicle speed.
- The authors have designed and implemented a new fusion module called the Local Attention-based Feature Fusion Module (LA-FFM). This unique strategy facilitates optimal interactions between each feature and its neighboring counterparts, which helps to enhance the level of multi-modal fusion. By using this module, this study are able to maintain the complexity of the model within manageable limits while improving the degree of multi-modal fusion.
- Through extensive experimentation and Comparison results, the authors validate the influence of input frames on prediction accuracy and conduct an in-depth analysis. The experimental results offer compelling evidence for the effectiveness of the MFE-SSNet.

2 Related Work

2.1 Optimal Driving

Optimal driving is a driving mode aimed at minimizing fuel consumption and reducing environmental impact while maintaining safety and legality. In recent years, research has been focused on vehicle cruise control, real-time steering angle prediction, speed optimization, and other related areas to achieve optimal driving.

Islam et al. [7] proposed a real-time steering angle prediction method based on digital maps, which can help drivers achieve optimal driving and reduce fuel consumption and emissions. This method employs route and road information from digital maps, combined with vehicle sensor data, to predict the upcoming turns and calculate the optimal turning speed. The method can also be adjusted based on real-time traffic conditions and road conditions to improve prediction accuracy. Pereira et al. [8] transformed the optimal driving problem into a multi-objective optimization problem, including minimizing travel time, fuel consumption, and emissions. This method uses real-time vehicle and environmental data for optimization, thereby generating the optimal speed curve. Lee et al. [9] proposed an optimal driving strategy based on real-time vehicle data and digital map information, including early deceleration, smooth driving, and avoiding sudden acceleration and braking. This method also considers the impact of road conditions and traffic situations to achieve optimal fuel efficiency and safety.

2.2 End-to-End Autonomous Driving

Deep neural networks (DNNs), particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are high-performance tools that can be used for processing images, videos, text, speech, and audio. They exhibit good performance in tasks that require building complex nonlinear mappings. In sharp contrast to existing frameworks, the development of DNNs has made it possible to control vehicles through an end-to-end approach, where raw sensor data, such as camera images or lidar point cloud data, is directly mapped to control commands, such as the angle of steering or the acceleration pedal.

In the early days, end-to-end autonomous driving methods were often achieved by configuring a simple network to directly map raw input data to control commands. Bojarski et al. [1] trained a CNN to directly map pixels from a single front-facing camera to steering commands and found it to be highly effective. Since then, many researchers have proposed new methods to improve the effectiveness of end-to-end self-driving approaches. For example, novel network architectures have been proposed to improve the effectiveness of end-to-end self-driving approaches [10, 11]. Chen et al. [12] proposed an auxiliary task network (ATN), where images are processed by a segmentation network and a DNN that operates in parallel to estimate optical flow. The segmentation and optical flow maps are then fed into CNN and LSTM, which output control commands. Xu et al. [13] proposed a fully convolutional LSTM (FCNLSTM) architecture that can learn from both segment loss and control loss. Unlike previous methods, self-driving tasks process the output of other tasks as input, and the tasks in FCN-LSTM share the same convolutional layer but have different fully connected layers. Wayve [14] achieved end-to-end autonomous driving by training an agent to learn the mapping from sensor data to vehicle control commands in a driving simulator. The process was divided into two stages: The first stage was trained using hand-designed rules to generate basic driving scenarios and behaviors as well as some basic driving skills; The second stage used a deep reinforcement learning algorithm (DDPG) to train the agent to perform end-to-end learning in the simulator, achieving comprehensive learning from perception to control.

2.3 Fusion Method

Multimodal fusion (MMF) refers to the integration of information from different modalities to achieve the more accurate and comprehensive understanding. The related work of MMF includes early and late fusion, deep learning methods, modality attention mechanisms, interactive MMF, ensemble learning methods, and weakly supervised MMF. In the context of autonomous driving, multimodal fusion is employed to fully leverage the complementary nature of speed and steering angle information to enhance the prediction performance of autonomous vehicles.

Liu et al. [15] proposed a novel shared-private framework that can leverage both textual and visual information for multimodal sentiment analysis. The framework employs a cross-modal prediction approach to learn shared representations across different modalities, while also maintaining private representations for each modality. Experimental results demonstrated the effectiveness of the proposed framework, which outperformed several state-of-the-art methods on benchmark datasets. Wang et al. [16] proposed an adversarial learning approach to learn joint representations of text and image data for click-through rate prediction. The proposed method employs a generative adversarial network to learn a mapping between textual and visual features, while also maximizing the mutual information between the joint representation and the click-through rate. Experimental results showed that the proposed method outperformed several baseline methods on a large-scale click-through rate prediction dataset. Jia et al. [17] proposed a multimodal fusion method, TFN, that leverages a tensor-based fusion approach to capture complex interactions between modalities. This method learns a joint representation of the input modalities by taking into account the interaction between modalities at the tensor level. Gao et al. [18] presented LMF, a method for multimodal fusion that utilizes low-rank decomposition to reduce the dimensionality of the multimodal input. This method aims to learn a joint representation of the input modalities by exploiting the low-rank structure of the multi-modal data. Dai et al. [19] proposed a novel approach (AFF) for feature fusion in computer vision tasks, which addresses the challenge of fusing features with inconsistent semantics and scales. In their paper, they proposed a multi-scale channel attention module to selectively weight and fuse features at different scales, and an iterative attentional feature fusion mechanism to alleviate the bottleneck issue that arises during the initial integration of feature maps.

In the context of autonomous driving, multimodal fusion has been utilized to combine speed and steering angle information from digital maps and vehicle sensors to predict the upcoming turns and calculate the optimal turning speed. This approach can help drivers achieve optimal driving, reduce fuel consumption, and emissions. The integration of multimodal information has also been shown to improve the robustness and adaptability of autonomous driving systems in various scenarios.

3 Method

3.1 Overview Architecture

The proposed Multi-Modal Fusion-Based End-to-End Steering Angle and Vehicle Speed Prediction Network (MFE-SSNet) comprises three main components: the three-stream feature extraction module, the Local Attention-based Feature Fusion Module (LA-FFM), and the final prediction module, as illustrated in Fig. 1.

MFE-SSNet takes road segment image data, along with steering angle and vehicle speed data from the previous 20 timestamps, as input. These data undergo initial processing by the three-stream feature extraction module, which leverages HRNet and LSTM layers to extract relevant features. The resulting feature vectors from the three modalities are subsequently fused using the LA-FFM, enabling the acquisition of comprehensive fused features. These fused features are ultimately employed for steering angle and vehicle speed prediction through the final prediction module.

Overall, the proposed approach effectively capitalizes on the complementary information provided by multiple modalities, thereby enhancing the accuracy and robustness of information detection. The experimental results validate the effectiveness and superiority of the method compared to other state-of-the-art methods.

3.2 Three-Stream Feature Extraction Module

As Shown in Fig. 1, the three-stream feature extraction module is structured into three primary streams: the image stream, the speed stream, and the steering stream. Each stream is uniquely equipped with a series of neural network components, including HRNet, LSTM layers, and MLP, to facilitate different methods of feature extraction tasks. Further comprehensive details regarding each component will be provided within the three-stream feature extraction module.

Image Stream The image stream is a significant component of the feature extraction module, accepting a sequence of temporal camera images as input. It employs the cutting-edge HRNet w48 network, boasting a 48-channel high-resolution structure to process the visual data. The HRNet w48 is meticulously crafted with down-sampling and up-sampling pathways within its blocks, facilitating a rich exchange of spatial information across feature maps of varying resolutions while upholding the integrity of high-resolution features.

Following the HRNet w48, the image data proceeds to an LSTM layer, architectured with a 256-unit hidden state size to robustly capture temporal dependencies and enhance the fidelity of feature extraction from the sequence of images. The LSTM's architecture, including its input, forget, output gates, and cell state, is essential for discriminating valuable feature information from noise. Thus, this emphasizes the long-term dependencies and pertinent patterns within the image data.

Speed and Steering Streams Addressing the speed and steering streams, the architecture harnesses a single LSTM layer, followed by a pair of fully connected (MLP) layers to process the temporal data. The LSTM is designed with 128 hidden units to precisely analyze the temporal dynamics inherent in both speed and steering data streams. Subsequent to the LSTM's meticulous processing, the data traverses through two MLP layers, designed to further refine the feature representation. The first MLP layer, encompassing 64 neurons, introduces a non-linear transformation using a

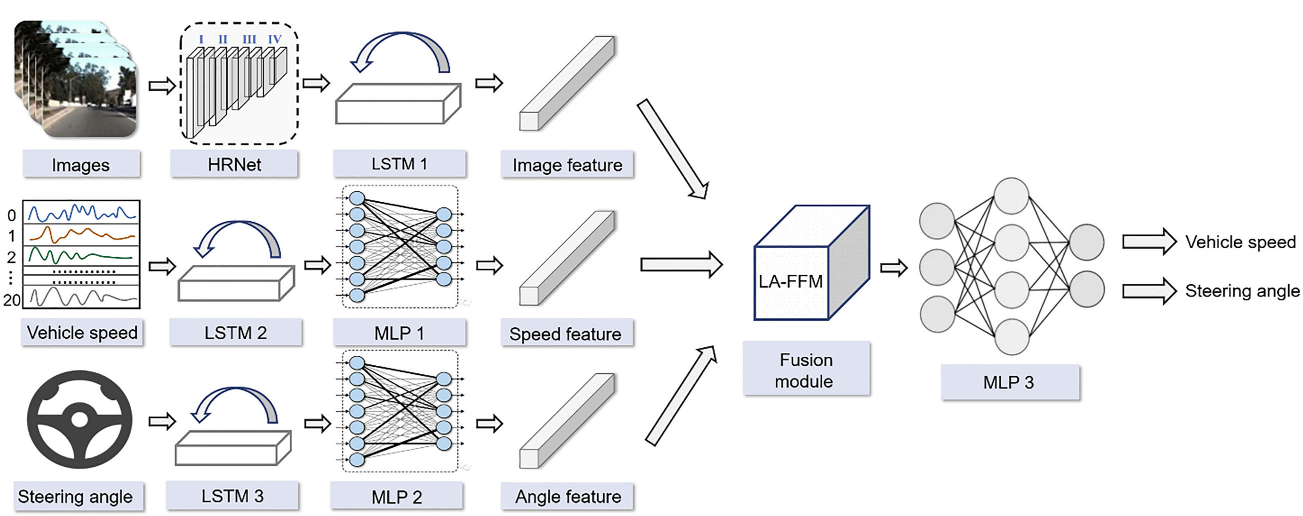

Fig. 1 The Workflow of MFE-SSNet

ReLU activation function, which is essential for capturing the complex relationships within the data. The output of this layer is then projected onto a second MLP layer, which performs a dimensionality reduction by mapping the 64-dimensional feature vector to a 32-dimensional space, while also employing a ReLU activation function to preserve the non-linearity within the network's processing.

These two MLP layers are integral in the stream's architecture, with their weight matrices and biases being iteratively optimized via back propagation, guided by a loss function that quantifies the prediction error. This optimization is vital for the model's ability to accurately predict the outcomes based on the fused features from the speed, steering, and image streams.

The fusion of the refined features from the speed and steering streams with the high-resolution features from the image stream results in a comprehensive feature vector that capitalizes on the complementary information from the various modalities.

In summary, the proposed approach significantly enhances the system's accuracy and robustness, which is substantiated by the experimental results highlighted in the following sections of the paper. The experimental evidence firmly establishes the superior performance of the method when compared to other state-of-the-art approaches, emphasizing the importance of detailing the HRNet, LSTM and MLP network architectures within complex multi-modal feature extraction frameworks.

3.3 Local Attention-Based Feature Fusion Module

This section introduces a novel feature fusion strategy called Local Attention Feature Fusion Module (LA-FFM). The module aims to enhance the fusion of feature vectors from different modalities by capturing the interdependencies within local channels.

The proposed fusion strategy consists of the following steps: First, images, vehicle speed, and steering angle features are extracted from the three-stream network and concatenated into a single feature map. Then, the LA-FFM module is used for local attention-based feature fusion. After these operations, the features are fed into the multi-layer perceptron (MLP) to obtain the final predicted steering angle and vehicle speed.

The innovation of the LA-FFM module lies in its direct interaction between each channel and its k nearest neighbors, thereby controlling model complexity and improving feature fusion. The LA-FFM module is defined as follows:

First, the three feature vectors obtained by the feature extraction module, namely the image feature vector, speed feature vector, and steering angle feature vector, are concatenated using the Concat function to form the initial fusion feature vector X.

$$X = \mathrm{Concat}(\mathrm{fimage},\ \mathrm{fspeed},\ \mathrm{fangle}) \tag{1}$$

where fimage denotes the image feature vector, fspeed denotes the speed feature vector, fangle denotes the steering angle feature vector, and Concat denotes the concatenate function used to form the initial fusion feature vector.

Next, the LA-FFM module is used for local attention-based feature fusion, defined as follows:

$$s = \sigma(\mathrm{Conv1D}(\mathrm{GAP}(X))) \tag{2}$$

$$Y = s \odot X \tag{3}$$

where Conv1D $(\cdot)$ denotes a 1D convolution with a kernel size of k in the channel domain, modeling local cross-channel interactions. GAP(X) denotes the channel-level global average pooling function, σ denotes the sigmoid function, s denotes the weight of the channel, and Y denotes the fused feature map.

After obtaining the initial fusion feature vector X, the authors perform channel-level global average pooling using GAP(X) to obtain the aggregated feature of the initial fusion vector. Then, a one-dimensional convolution is performed to capture the interdependencies between local channels and obtain a new feature map. This paper employs the sigmoid function to convert these features into weights between 0 and 1. The original feature map X is multiplied by the obtained weight s to obtain Y, thereby achieving feature enhancement.

Upon obtaining the aggregated features acquired by Global Average Pooling (GAP), the authors employ the Local Adaptive Feature Fusion Module (LA-FFM), as depicted in Fig. 2. The LA-FFM is designed to optimize the channel attention mechanism within CNN models. This module leverages the aggregated features from GAP and

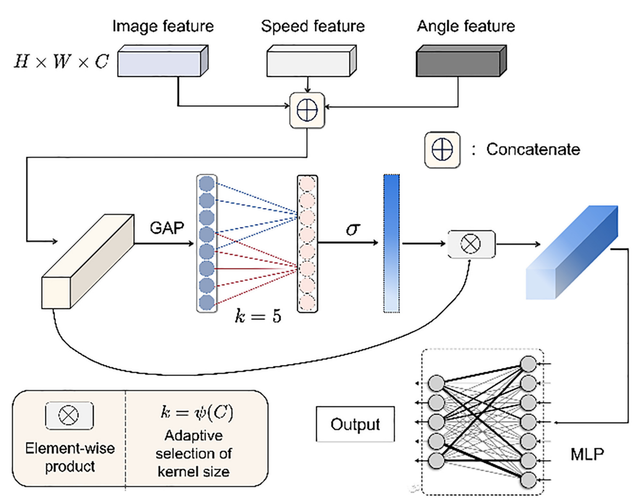

Fig. 2 Illustration of the local attention-based feature fusion module (LAFFM)

generates channel weights by performing swift 1D convolutions, thereby enhancing the interaction among features. In the process of fixing the kernel size k in all convolution blocks, the experimental investigations revealed that LA-FFM performs optimally when k is either 9 or 5. Consequently, a kernel size of 5 is selected for the one-dimensional convolution operation to strike an effective balance between computational efficiency and the capacity for capturing pertinent local cross-channel interactions.

The kernel size, k, of the 1D convolution is a critical parameter that defines the extent of local cross-channel interactions. This paper has discarded the strategy of dimensionality reduction to prevent potential information loss and has adopted an adaptive method to determine the optimal kernel size k, ensuring effective cross-channel interactions while maintaining low model complexity.

As illustrated in Fig. 2, the LA-FFM module, with a fixed kernel size, facilitates effective local cross-channel interactions without reducing dimensionality by considering each channel and its k neighbors. To adaptively determine k without relying on cross-validation, the authors assume that the kernel size k is directly proportional to the channel dimension C, and the following mapping is developed:

$$C = \varphi(k) \tag{4}$$

While the simplest mapping is a linear function:

$$\varphi(k) = \gamma \cdot k + b \tag{5}$$

Considering that the channel dimension C is typically set to a power of 3, this paper extends the linear mapping to a non-linear function:

$$C = \varphi(k) = 3\gamma \cdot k + b \tag{6}$$

To adaptively determine the kernel size k, the authors employ the following method:

$$k = \psi(C) = \left| \frac{\log_3(C) - b}{\gamma} \right|_{\text{odd}} \tag{7}$$

where $|\cdot|_{\text{odd}}$ denotes the operation of rounding to the nearest odd integer. For all the experiments, the authors set γ and b to 2 and 1, respectively. Through the mapping ψ, channels with higher dimensions engage in longer-range interactions, while those with lower dimensions engage in shorter-range interactions.

After these operations, the fused features are used as the input layer of an MLP model. Each input feature is multiplied by a weight vector and passed through an activation function for nonlinear transformation, mapping the input data to the next layer's neurons. The output layer contains two neurons, representing the predicted vehicle speed and steering angle, respectively, to obtain the final prediction results. The MLP model can be represented by the following equations:

$$H = \varnothing(XW_h) + b_h \tag{8}$$

$$O = HW_o + b_o \tag{9}$$

Here, H represents the output of the first layer, $\varnothing(\cdot)$ denotes the activation function, XW_h represents the first layer weight vector, and b_h represents the first layer bias vector. O represents the output of the second layer, W_o represents the second layer weight vector, and b_o represents the second layer bias vector. The outputs of the first and second layers represent the predicted values of speed and steering angle, respectively.

The innovation lies in the application of the LA-FFM module to the fusion of multimodal information such as images and speed for autonomous driving to accurately predict driving mode, vehicle speed, and steering angle. The LA-FFM module performs strongly in sparse scenes, especially when direct interaction between channels is limited. The experimental results show that the method using the LA-FFM module is significantly better than traditional fusion strategies in terms of accuracy and robustness.

4 Dataset and Evaluation Matrices

4.1 Dataset

The Udacity Challenge 2 dataset has been selected for this project. This challenge was initiated by Udacity's Self-Driving Car Initiative as the second challenge, with the goal of replicating the achievements of Nvidia's DAVE-2 deep learning system using their own design and implementation of a convolutional neural network. The DAVE-2 system is capable of teaching a car to drive solely using cameras and deep learning. It can drive in various weather conditions, avoid obstacles, and even handle off-road scenarios. The main task of the challenge is to use convolutional neural networks and deep learning to predict the appropriate steering angle from the image frames captured by the camera mounted on the car's windshield. The dataset contains six video clips, with a total duration of around 20 min. Speed values, steering angles and video streams from three front view cameras are recorded.

4.2 Data Pre-processing

A series of preprocessing steps was performed to ensure the quality and suitability of the dataset. Firstly, the velocity and steering angle data for each timestamp were stored in separate text files for easy processing and analysis. During the preprocessing phase, it was observed that low-speed data could introduce unnecessary interference in subsequent analyses. Therefore, it was decided to exclude data points with velocities below 4 m/s and reevaluate the dataset's validity. This decision was motivated by the fact that at very low vehicle speeds, there is a significant variation in steering angles, which could introduce noise and potentially impact the training and performance evaluation of the models. To divide the dataset, a 5:1 ratio was employed. Specifically, the third CSV file provided by the Udacity Challenge 2 was designated as the test set, while the remaining five CSV files were used as the training set. This partitioning scheme ensured that sufficient data was available for model training and that an independent dataset was available for assessing the model's generalization performance.

Through these preprocessing steps and dataset partitioning, the authors curated a filtered and optimized dataset that supports the subsequent research and experiments. This dataset preparation provides a reliable foundation for in-depth analysis and understanding of the relationship between vehicle speed and steering angle, thereby enhancing the applicability of the research objectives.

4.3 Loss Functions

In the task, the authors independently predict driving speed and steering angle. For this, two independent Root Mean Square Error (RMSE) loss functions are defined to evaluate the accuracy of speed and steering angle predictions, respectively. RMSE can reflect the average magnitude of the deviation between predicted and actual values. In the prediction of driving speed and steering angle, significant deviations should be regarded with caution. The specific loss functions are defined as follows: For speed, the loss function is defined as:

$$L_{\text{speed}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{speed},i} - \hat{y}_{\text{speed},i}\right)^2} \tag{10}$$

For the steering angle, the loss function is defined as:

$$L_{\text{angle}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{angle},i} - \hat{y}_{\text{angle},i}\right)^2} \tag{11}$$

where $y_{\text{speed},i}$, $y_{\text{angle},i}$, $\widehat{y}_{\text{speed},i}$, and $\widehat{y}_{\text{angle},i}$ represent the actual speed, actual steering angle, predicted speed, and predicted steering angle of the i-th frame image, respectively.

The test set contains a total of n frames. A smaller loss function value indicates a higher prediction accuracy of the model.

Lastly, this paper adds the loss functions of speed and steering angle to form the total loss function. Specifically, the total loss function is calculated by the sum of the loss function of speed and the loss function of the steering angle, as follows:

$$L_{\text{total}} = L_{\text{speed}} + L_{\text{angle}} \tag{12}$$

This design allows the loss function to consider the accuracy of driving speed and steering angle predictions simultaneously, thus more comprehensively evaluating the performance of the model.

5 Experiment

5.1 Training Details

The experiment was trained on the Udacity dataset. In this paper, a subset of the Udacity dataset Udacity Challenge-II is adopted for the experimental purpose. The Udacity Challenge-II dataset contains 33,808 frames for model training and 5614 frames for model testing, in which vehicle speed, steering angle, and video streams from three front view cameras are recorded. The resolution of images in Udacity is 480 * 640. The training experiments are conducted on a server with 1 Geforce RTX 3090 GPU. All code is written in the Pytorch framework.

The following is some crucial parameters for re-implementing the method. The input images are resized to the size of 224 * 224. Adam optimizer is used with the learning rate of 1e−4. This paper randomly draws 5% of the training data for validating models and always memorizes the best model on this validation set. Training a model requires about 1–2 days over a single GPU.

5.2 Evaluation Results on Udacity Dataset

To explore the effect of input history frames on the model's prediction of steering angle and speed, experiments were conducted on data streams with history frames from 2 to 20. The authors utilize the widely recognized Udacity Challenge-II dataset, which mirrors real-world dynamic driving scenarios, to validate the efficiency and reliability of the proposed driving model. It should be noted that the experimental results represent the average of five trials, rather than selecting the best result from each individual trial.

The results of the effect of different input history frames on the predicted speed of the model are illustrated in Table 1. In this case, frames denote the number of input history frames, t denotes the current frame, and $t+n$ denotes the predicted speed value for the next n frames. Taking the input history of 2 frames as an example, the prediction of speed is 0.377 at 11 predicted frames. As the number of predicted frames increases, the model's effectiveness in predicting

Table 1 Speed experimental results of the proposed model on the Udacity Challenge-II dataset

Frames	$t+11$	$t+12$	$t+13$	$t+14$	$t+15$	$t+16$	$t+17$	$t+18$	$t+19$	$t+20$
2	0.377	0.401	0.426	0.435	0.462	0.477	0.495	0.528	0.542	0.561
4	0.349	0.373	0.397	0.419	0.441	0.466	0.488	0.511	0.531	0.554
6	0.358	0.388	0.424	0.427	0.456	0.469	0.491	0.521	0.537	0.558
8	0.359	0.377	0.431	0.431	0.459	0.472	0.494	0.527	0.539	0.560
10	0.361	0.378	0.436	0.437	0.466	0.481	0.492	0.528	0.535	0.564
12	0.362	0.365	0.423	0.424	0.452	0.483	0.491	0.521	0.536	0.567
14	0.367	0.371	0.413	0.426	0.447	0.486	0.481	0.512	0.524	0.565
16	0.369	0.374	0.416	0.419	0.439	0.488	0.482	0.511	0.527	0.569
18	0.381	0.379	0.428	0.436	0.457	0.489	0.502	0.532	0.547	0.572
20	0.402	0.423	0.445	0.466	0.490	0.511	0.531	0.546	0.579	0.586

speed gradually decreases. At 20 frames, the model's performance is at its worst, with a prediction of 0.561. For the prediction of the same future frame, taking the prediction of frame $t+11$ as an example, when the number of input frames is 4, the model predicts the speed result of 0.349, which is better than all other input history frames. As the number of input history frames increases, the model's prediction speed performance decreases and is worst at 20 input frames, with a prediction performance of 0.402. For this phenomenon, it is believed that inputting a certain number of history frames can help the model achieve better prediction results, but inputting an excessively large number of history frames may result in the redundancy of information and lead to a drop in the prediction performance of the model. It is worth noting that the experimental results illustrated in Table 1 show relative consistency across frames, which indicates that the model remains stable in the face of different driving conditions. However, it is evident that the root mean square error (RMSE) values fluctuate slightly across the different prediction timestamps. This is to be expected given the complexity and variability of real driving scenarios. Nevertheless, these deviations are moderate and fall within acceptable limits, emphasizing the robustness of the model in speed prediction.

Meanwhile, Table 2 shows the effect of different input history frames on the model's predicted steering angle. Similarly, with an input history of 2 frames, the model predicts the steering angle with a result of 0.099 at frame $t+11$. As the number of frames predicted continues to increase, the model's effectiveness in predicting the steering angle decreases, with the worst result being 0.143 at frame $t+20$. When predicting the future frame number constant, for example, with a prediction frame of $t+11$, the model predicts the steering angle at 0.095 with an input history of 4, which is represents the most accurate prediction performance among all history frames. As the number of input frames increases, the model's performance in predicting the steering angle begins to fluctuate but is lower than the performance at frame 4. The model's performance in predicting steering angle is worst at 0.108 at 16 input frames. It is hypothesized that this phenomenon may be due to the fact that steering angle changes often occur within a short period of time. If too many frames are input, this may cause interference in the prediction process, resulting in poor and fluctuating prediction accuracy. As with the performance demonstrated in Table 1, the model also maintains a strong consistency in predicting steering angles, demonstrating its adaptability in dealing with a wide range of driving conditions. While there are variations in the results, these fluctuations are within tolerable limits. Thus, these results showcase the model's

Table 2 Steer experimental results of the proposed model on the Udacity Challenge-II dataset

Frames	$t+11$	$t+12$	$t+13$	$t+14$	$t+15$	$t+16$	$t+17$	$t+18$	$t+19$	$t+20$
2	0.099	0.103	0.108	0.115	0.120	0.125	0.128	0.137	0.139	0.143
4	0.095	0.100	0.105	0.111	0.116	0.121	0.125	0.130	0.135	0.139
6	0.105	0.103	0.107	0.112	0.118	0.128	0.133	0.146	0.149	0.159
8	0.102	0.104	0.106	0.110	0.117	0.127	0.131	0.138	0.148	0.151
10	0.101	0.107	0.108	0.111	0.119	0.126	0.132	0.136	0.145	0.149
12	0.103	0.108	0.113	0.119	0.124	0.127	0.134	0.145	0.146	0.154
14	0.099	0.105	0.107	0.114	0.121	0.125	0.131	0.139	0.141	0.149
16	0.108	0.103	0.109	0.115	0.118	0.129	0.137	0.144	0.150	0.163
18	0.100	0.102	0.107	0.118	0.121	0.125	0.131	0.139	0.144	0.146
20	0.104	0.102	0.108	0.114	0.119	0.126	0.129	0.134	0.139	0.149

adept skill in accurately predicting steering angles even in the face of changing conditions.

In conclusion, the experimental results compellingly substantiate the superior performance of the proposed driving model on the Udacity Challenge-II dataset, especially when compared with other contemporary methods. Despite confronting the challenges embodied in the dataset reflective of actual driving scenarios, it excels in predicting both speed and steering angle. The model's capability in this environment not only emphasizes its theoretical validity but also highlights its potential applicability in practical autonomous driving systems.

Table 3 Experimental results of steering angle prediction on Udacity Dataset

Method	RMSE
3D CNN + LSTM [20]	2.7167
3D ResNet + LSTM [20]	2.4899
FM-Net [20]	2.3549
MH-IND-FC [21]	1.8970
MH-SIM-FC [21]	1.8970
MH-SIM-LSTM [21]	1.3940
MSINet [22]	0.0491
MFE-SSNet (the authors)	0.0300

5.3 Qualitative Results

In this section, the authors performed qualitative analyses on the Multi-Modal Fusion End-to-End Steering Angle and Speed Prediction Network (MFE-SSNet) proposed in this study, using two different dimensions. The first analysis focused on the predicted results of the steering angle and speed for different prediction frames, with 0–20 input frames. This paper selected 11, 16, and 20 prediction frames for analysis and used different input frames as the horizontal axis. The resulting RMSE is the vertical axis. The second analysis evaluated the overall prediction performance for different prediction frames, with the steering angle and speed prediction results for 11 to 20 prediction frames plotted on the horizontal axis.

To provide a comprehensive and intuitive evaluation of the prediction analysis, the visualization methods for qualitative analysis were utilized.

5.4 Comparison with State-of-the-Art Models

This section presents a comparison of the performance of the method proposed in this paper with other advanced algorithms on the Udacity dataset. This paper's focus is on evaluating the prediction of two types of information: steering angle and speed. However, it is crucial to acknowledge that there are variations in the number of future frames selected for prediction by each method. To ensure a fair experimental comparison, the authors have chosen the best prediction performance from each method and compared it with the best performance achieved by the proposed model. It is determined that the optimal prediction performance of the proposed model is obtained when using four input historical frames and predicting one future frame. The detailed analysis is presented in Sect. 5.4.

Table 3 presents the comparison of steering angle prediction performance among different models, including FM-Net proposed by Hou et al. [20], MH-SIM-LSTM proposed by Kosman et al. [21], and MSINet proposed by Wu et al. [22]. The proposed MFE-SSNet demonstrates superior performance compared to these competing methods. According to the root mean square error (RMSE) evaluation shown in this table, the model achieves an RMSE of 0.0300, which is significantly better than the 2.3549 recorded by FM-Net. When compared to the method proposed by Kosman, the method outperforms all three of their proposed models and substantially surpasses their best-performing method. While the RMSE value may not show a significant improvement compared to the MSINet model, MFE-SSNet offers several advantages over the MSINet model, with the most prominent being its input. Unlike MSINet, which only considers image information, the MFE-SSNet incorporates additional data such as vehicle speed and steering angle. By integrating multimodal information into the input, it is believed that the model can capture more relevant features and thereby achieve better prediction results. The combination of image and vehicle speed data allows the model to learn from diverse sources, leading to enhanced predictive capabilities. The comparison results with other advanced models exhibit a significant improvement in predicting steering angle using the model. This improvement is particularly crucial for high-precision autonomous driving systems, where accurate steering angle prediction plays a key role.

As most end-to-end autonomous driving models primarily focus on steering angle prediction rather than speed, this paper conducted a comparison with some new models, including MS-MSNet proposed by Kosman et al. [23] and DP-Net proposed by Xiong et al. [24], to evaluate the speed prediction performance. As depicted in Table 4, the model achieves a prediction performance of 0.0840 in terms of speed, outperforming the MH-SIM-LSTM model significantly on the same Udacity dataset. This improvement can be attributed to the detailed treatment of multimodal information in the MFE-SSNet model compared to MH-SIM-LSTM. While the MH-SIM-LSTM model also incorporates multiple-modal inputs, it lacks sufficient detail in handling these inputs, merely concatenating them. In contrast, the model incorporates a multimodal feature fusion module called LA-FFM, which allows for

Table 4 Experimental results of steering angle prediction on Udacity Dataset

Method	RMSE
MS-MSNet [23]	2.1480
DP-Net [24]	1.4211
MH-IND-FC [21]	2.6730
MH-SIM-FC [21]	1.8990
MH-SIM-LSTM [21]	0.3130
MFE-SSNet (the authors)	0.0840

comprehensive information fusion across different modalities. This facilitates the mutual assistance of modalities, enabling the extraction of more valuable feature information. Therefore, the proposed model effectively improves the performance of existing models in predicting vehicle speed, enabling more accurate speed prediction in the real world. This is an important aspect of maintaining stable and safe driving conditions.

In the realm of steering angle prediction, the MFE-SSNet model achieves a significant performance improvement over the widely accepted MSINet model. The method accurately predicts the steering angle across various complex and changing driving scenarios, which include, but are not limited to, narrow roads, vehicle turning, and various weather conditions. Moreover, the model significantly surpasses MSINet in stability and accuracy when handling large-angle turns. This performance improvement can be attributed to the model's ability to consider a wide range of real-world driving scenarios. In terms of speed prediction, the MFE-SSNet model has a significant advantage over the comparison models, mainly due to the fact that the model performs well when dealing with complex situations such as high-speed driving and sudden deceleration. In addition, the model maintains a higher accuracy when dealing with situations such as traffic congestion and lane changes. These small improvements emphasize the strengths of the model when dealing with detailed situations and contribute to stable, safe, and efficient driving.

In conclusion, the MFE-SSNet model demonstrates substantial improvements over other approaches, making it capable of achieving significant advancements in complex driving environments. These improvements not only provide theoretical benefits but also hold the promise of delivering superior performance and an enhanced driving experience in real-world applications, particularly in high-precision and high-quality autonomous driving systems. In future research, the authors will continue refining the model and conducting additional field tests to further validate the practicality and effectiveness of MFE-SSNet. This work aims to strengthen the robustness and reliability of the model, ensuring its viability for real-world deployment.

5.4.1 Predicted Results of Speed and Steering Angle at Different Prediction Frames with 0–20 Input Frames

The Fig. 3(a)–(f) illustrate the performance of the MFE-SSNet model in predicting steering angles and vehicle speeds. Figure 3(a) through 3(c) depict the results for speed prediction, while Fig. 3(d) through 3(f) display the visualization results for steering angle prediction at different prediction frames with 0–20 input frames.

As shown in Fig. 3(a)–(c), the prediction results for the first two input frames exhibit a relatively high error overall. Despite this, the prediction error for vehicle speed decreases gradually as the number of input frames increases from 2 to 4. This reduction in error contributes to lower RMSE values, particularly as the optimal prediction occurs with four input frames, which is presumably due to the completion of the first set of introduced motions. Consequently, even as the number of input frames continues to increase beyond this point, the RMSE remains relatively low, indicating a maintained level of prediction accuracy despite the introduction of potential noisy motions.

Regarding steering angle prediction, Fig. 3(d)–(f) show that the trends in steering angle and speed prediction are generally similar. The model achieves the lowest error and thus low RMSE values when four frames are used as input. Even with small fluctuations in prediction performance within the range of 6 to 20 frames, the overall trend remains stable, which contributes to maintaining low RMSE values.

These results show that the model can accurately predict steering angles and vehicle speed under various road and driving conditions, highlighting its robustness and reliability.

5.4.2 Predicted Results of Steering Angle and Vehicle Speed for 11 to 20 Frames with All Input Frames

For the investigation of the optimal accuracy among the predicted frames when the input frames remain unchanged, a qualitative analysis was employed, as depicted in Fig. 4(a), (b). In the box plot, the purple circles represent all data points when predicting speed, while the red circles represent all data points when predicting steering angle.

As illustrated in Fig. 4(a), in the case of vehicle speed prediction, the error values among different input frames remain relatively minimal as the prediction frame progresses from 11 to 20, suggesting a stable model. This minimal variation in error values results in a low RMSE index of 0.0840, as presented in Table 4.

In the case of steering angle prediction, as illustrated in Fig. 4(b), the range of RMSE values exhibits a relatively consistent pattern, contributing to a low RMSE index of 0.030, as displayed in Table 3. The consistency in the RMSE values, despite the increase in prediction frames,

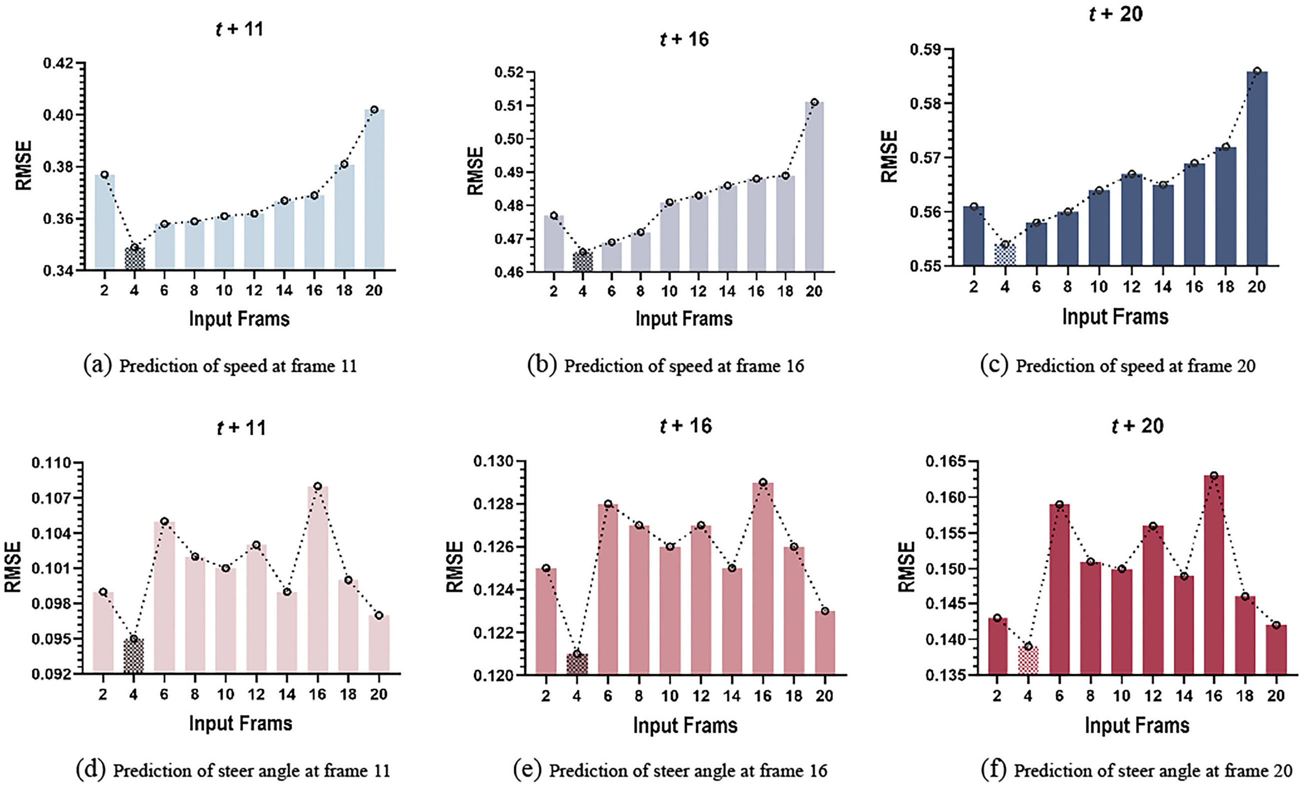

Fig. 3 Predicted results of speed and steering angle on the Udacity dataset. **a**–**c** Show the speed predictions at frames 11, 16, and 20 respectively. **d**–**f** Show the steering angle predictions at the same frames

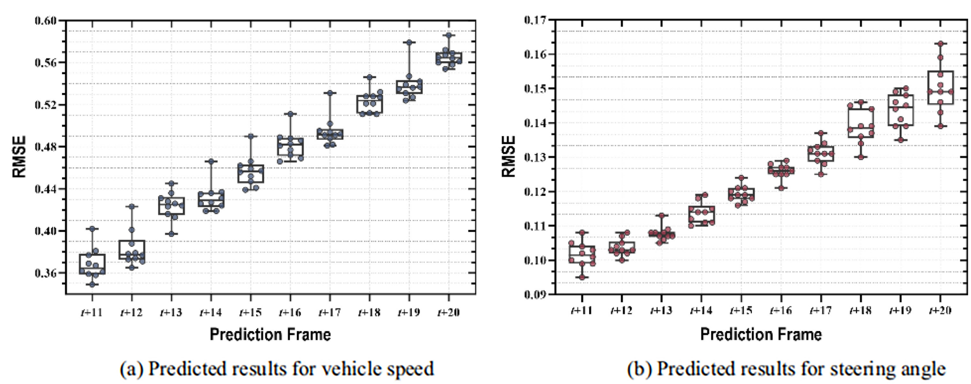

Fig. 4 Speed and steering angle prediction result for 11 to 20 frames. **a** Shows the statistical results of all input frames for speed prediction. **b** Shows the statistical results of all input frames for steer angle prediction

is indicative of the model's ability to maintain accurate predictions.

These results highlight the model's ability to generate accurate predictions of steering angles and vehicle speeds despite variations in road and driving conditions, thereby attaining low RMSE values. As such, the MFE-SSNet model has demonstrated its robust and effective predictive performance in real-world driving scenarios, highlighting its potential for practical applications in autonomous driving systems.

In conclusion, the results illustrate the outstanding performance of the proposed MFE-SSNet model in predicting steering angles and vehicle speeds. Consequently, the proposed end-to-end multimodal fusion-based steering angle and vehicle speed prediction network (MFE-SSNet) demonstrates robust and effective predictive performance in real-world driving scenarios, highlighting its potential for practical applications in autonomous driving systems.

Table 5 Results of the ablation study on different fusion modules

Add	Basic fusion	LA-FFM	RMSE (Speed)	RMSE (Angel)
✓			0.118	0.090
	✓		0.102	0.060
		✓	0.084	0.030

Table 6 Results of the ablation study on network input architectures

Image	Speed	Angel	RMSE (Speed)	RMSE (Angel)
✓			0.156	0.070
	✓		0.124	0.150
		✓	0.161	0.050
✓	✓		0.121	0.060
✓		✓	0.129	0.050
✓	✓	✓	0.084	0.030

5.5 Ablation Studies

The comprehensive experiments were conducted to verify the effectiveness of each key component of the model, as shown in Tables 5 and 6. This paper explored two key components, namely the local attention-based feature fusion module (LA-FFM) and the three-stream network architecture. In the ablation experimental part, the authors all use 4 frames as input to the model, while judging the change in performance by analyzing the accuracy of predicting the vehicle speed and turning angle in the first moment of the future.

5.5.1 Ablation Experiments on the Components of LA-FFM

In order to gain a deeper understanding of the impact of the Local Attention-based Feature Fusion Module (LA-FFM) on predicting vehicle speed and steering angle, this paper designed and conducted a series of ablation experiments.

Firstly, LA-FFM is employed to fuse information from different modalities. This is also the optimal result the experiments achieved, with an RMSE of 0.084 for vehicle speed and 0.03 for steering angle. The performance of this fusion module serves as a baseline for comparison with subsequent ablation experiments.

Next, a model was examined which did not employ any special fusion methods. In this variant, the authors did not employ any special fusion strategy, but instead did a simple summation of the feature vectors of the different modes. As can be seen in Table 5, the RMSE metric achieved by the model using summation as the fusion strategy is higher than the baseline of 0.034, with a significant decrease in the effect, thus confirming the effectiveness of LA-FFM in the prediction task.

Furthermore, a model version was considered that uses a basic fusion mechanism, such as simple averaging or weighted averaging, to fuse modality information. While this basic fusion mechanism performed better than the modality-independent version, it still fell short of the baseline model's performance.

This phenomenon is also expected, given that the simple fusion approach cannot fully exploit the interrelationship of different modes, which leads to limited effective features for the final prediction, thus resulting in an increased RMSE error.

The detailed results of these experiments can be found in Table 5. The experimental findings strongly suggest the superiority of the attention-based modality fusion strategy in the prediction task compared to other strategies. By allocating different weights to each modality, the attention mechanism can effectively fuse modality information, resulting in more accurate predictions. Therefore, the research provides compelling evidence for the use of the attention mechanism as a modality fusion strategy in related work.

5.5.2 Ablation Experiments on Network Input Architectures

In the second set of ablation studies, the authors investigated the effect of different network input architectures on the prediction task. Six different input configurations were tested, each represented by a unique model variant.

The "Image Model" uses only image data as input. Although this variant is able to make reasonable predictions, its performance is significantly worse than the full model. Also, as can be seen in Table 6, The RMSE of steering angle has been as low as 0.07, but the RMSE of vehicle speed has reached 0.156, which is 0.072 higher than the optimal effect. The time-series image data responds better to the turn angle, but less significantly to the vehicle speed. This phenomenon can be interpreted as the network can better learn the angle changes in the time-series images, but the representation for the speed is relatively weaker. However, the comparison of the results shows the importance of non-image data, such as speed and steering angle, in making accurate predictions.

The "Speed Model" and "Angle Model" use only vehicle speed and steering angle data, respectively. Firstly, when only the speed is used as input, the accuracy of speed prediction is significantly improved compared to using only image, and RMSE was reduced by 0.032. However, the prediction of steering angle is poor, and RMSE was improved by 0.08. Secondly, when using only the steering angle as input, there is some improvement in the prediction of the steering angle, but the prediction of the speed also decreases. For this phenomenon, the authors conjecture that one-dimensional data express limited features, and it is more difficult to represent the steering angle only by the speed or to represent the speed by the steering angle.

The "Image-Speed model" and the "Image-Angle model" then use only speed and steering angle data or image and speed as inputs. It can be found that both models have a greater improvement in accuracy when predicting either vehicle speed or steering angle. It is hypothesized that the angle change of the image and the degree of blurring reflect the change of steering angle and vehicle speed, and have more common features with vehicle speed and turning angle. Therefore, when the vehicle speed or steering angle is fused with the image for prediction, the prediction accuracy is significantly improved.

Finally, the “Full model” uses a three-stream architecture that takes as input image data, vehicle data and steering angle data. As expected, this variant outperformed the other five variants, achieving the lowest RMSE on test sets. The RMSE for vehicle speed and steering angle reached 0.084 and 0.03, respectively.

In scrutinizing the results of the ablation studies presented in Table 6, the authors discerned the profound impact of different network input architectures on the predictive task. Six distinct input configurations were examined, each represented by a unique model variant.

Firstly, the ‘Image Model’ solely relied on image data as input. While this variant was capable of generating sensible predictions, its performance was decidedly inferior compared to the complete model. This underlines the significance of non-image data, such as speed and steering angle, in making accurate predictions. Subsequently, the “Speed Model” and “Angle Model” utilized only speed and steering angle data, respectively, in the absence of any image information. These variants performed subpar to the Image Model, further solidifying the importance of image data in the task. Following, the ‘Speed-Angle Model’ and ‘Image-Speed Model’ employed either speed and steering angle data or image and speed as input, without any image information or steering angle information, respectively. These two variants outperformed their predecessors, reinforcing the importance of an integrated network. Finally, the “Full Model” leverages a tri-stream architecture that assimilates both image data and speed-angle data as inputs. As anticipated, this configuration outperformed the other five, attaining the lowest RMSE on both the training and test sets.

These results, detailed in Table 6, strongly suggest that a three-stream input architecture, which incorporates both image and speed-angle data, is the most effective for this prediction task. This highlights the importance of using multi-modal data.

6 Conclusion

This study introduces a pioneering Multi-Modal Fusion-Based End-to-End Steering Angle and Vehicle Speed Prediction Network (MFE-SSNet), utilizing a three-stream architecture that efficaciously leverages the supplementary information across multiple modalities. Additionally, an innovative feature fusion strategy termed was proposed as Local Attention-based Feature Fusion Module (LA-FFM). This module augments the fusion of varying modal feature vectors by recognizing the interdependencies in local channels. In comparison to the currently leading model algorithms, the MFE-SSNet procures more precise results on the publicly accessible Udacity dataset, thereby substantiating the effectiveness of MFE-SSNet. This research paves the way for a novel and efficient solution to the problems of multi-modal fusion and end-to-end driving prediction, with potential widespread applications in autonomous driving technology.

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest.

References

1. Bojarski, M., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
2. Abbas, M.A., Milman, R., Eklund, J.M.: Obstacle avoidance in real time with nonlinear model predictive control of autonomous vehicles. Can. J. Electr. Comput. Eng. **40**(1), 12–22 (2017)
3. Wu, T., Luo, A., Huang, R., et al.: End-to-end driving model for steering control of autonomous vehicles with future spatiotemporal features. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 950–955. IEEE (2019)
4. Wang, T., Luo, Y., Liu, J., et al.: End-to-end self-driving approach independent of irrelevant roadside objects with auto-encoder. IEEE Trans. Intell. Transp. Syst. **23**(1), 641–650 (2020)
5. Wang, Q., Wu, B., Zhu, P., et al.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)

6. Liu, Z., Huang, T., Li, B., et al.: Epnet++: cascade bi-directional fusion for multimodal 3D object detection. IEEE Trans. Pattern Anal. Mach. Intell.Intell. (2022). https://doi.org/10.1109/TPAMI.2022.3228806
7. Islam, S.M., et al.: Real-time steering angle prediction for optimal driving using digital map. IEEE Trans. Intell. Transp. Syst. **19**(9), 2897–2906 (2018)
8. Pereira, F.L., et al.: Multi-objective optimization approach for eco-driving assistance systems. IEEE Trans. Intell. Transp. Syst. **14**(1), 376–387 (2013)
9. Lee, J.: Real-time eco-driving strategy for improving fuel efficiency and safety using vehicle-to-infrastructure communication. IEEE Trans. Intell. Transp. Syst. **16**(1), 94–103 (2015)
10. Huang, X., et al.: Multi-modal prediction for autonomous driving using deep regression networks. IEEE Trans. Intell. Transp. Syst. **19**(3), 869–878 (2018)
11. Hou, Y., et al.: Learning a hierarchical driving policy using convolutional neural networks. IEEE Trans. Intell. Transp. Syst. **21**(1), 115–124 (2020)
12. Chen, X., et al.: End-to-end learning for lane keeping of self-driving cars. IEEE Intell. Transp. Syst. Mag. **7**(4), 42–52 (2015)
13. Xu, H., et al.: Learning to drive a high-dimensional discrete action space for autonomous driving. In: Proceedings of the 2017 Conference on Robot Learning, pp. 484–493 (2017)
14. Bansal, M., et al.: Chauffeurnet: learning to drive by imitating the best and synthesizing the worst. In: Proceedings of the 2018 Conference on Robot Learning, pp. 464–475 (2018)
15. Liu, J., Cai, D., Zhu, L., Liu, Y.: A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), pp. 4780–4789 (2021)
16. Wang, J., Deng, Z., Li, Y., Wang, Y.: Adversarial multimodal representation learning for click-through rate prediction. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021), pp. 2349–2353 (2021)
17. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: TFN: a deep network for multimodal fusion with high-level tensor representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 826–835 (2019)
18. Gao, H., Wang, X.-Y., Ji, R., Liu, W., Tao, D.: Low-rank multimodal fusion for multimedia analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1311–1320 (2019)
19. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3559–3568 (2021). https://doi.org/10.1109/WACV48630.2021.00360
20. Hou, Y., Ma, Z., Liu, C., Loy, C.C.: Learning to steer by mimicking features from heterogeneous auxiliary networks. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33018433. https://doi.org/10.1609/aaai.v33i01.33018433
21. Kosman, E., Castro, D.D.: Vision-guided forecasting—visual context for multihorizon time series forecasting (2021)
22. Wu, T., Luo, A., Huang, R., Cheng, H., Zhao, Y.: End-to-end driving model for steering control of autonomous vehicles with future spatiotemporal features. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 950–955 (2019). https://doi.org/10.1109/IROS40897.2019.8968453
23. Kosman, E., Castro, D.D.: Vision-guided forecasting—visual context for multihorizon time series forecasting. arXiv:2107.12674 (2021)
24. Xiong, H., Liu, H., Ma, J., Pan, Y., Zhang, R.: An NN-based double parallel longitudinal and lateral driving strategy for self-driving transport vehicles in structured road scenarios. Sustainability **13**(8), 4531 (2021). https://doi.org/10.3390/su13084531