IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 16, Issue 3, March 2026, ||Series -1|| PP 33-41

Vehicle Intention Detection In Long-Tail Scenarios Via Spatial-Temporal Fusion Of C2F_EMA And LSTM

Jun-Shen Chen; Yu-Hao Feng; Si-Dan Chen; Si-Yu Yin; Jun-Jie Zhang; Qing-Ming Zhang
Jiangsu University Of Technology, Changzhou, Jiangsu, 213001, China

Abstract
This paper aims to explore a vehicle intention and driving safety monitoring method based on improved YOLO architecture and Long Short-Term Memory (LSTM) network. By integrating the C2F_EMA module with an LSTM sequential network, the proposed method addresses the limitations of single-frame vision in "long-tail" scenarios (e.g., night-time glare). It significantly enhances the detection accuracy and temporal robustness of vehicle intentions, such as sudden braking and lane changes, thereby providing comprehensive support for driving safety. The research content covers the spatial information recognition based on C2F_EMA feature enhancement, dataset construction involving both simulated environments and real-world scenarios, and the stable output of LSTM temporal intention based on edge computing. Experiments verified the effectiveness and superiority of the proposed end-to-end perception method on the Jetson Orin Nano platform, which provides theoretical basis and technical support for ADAS safe driving detection methods.
Keywords: Intelligent assisted driving; Object detection; C2F_EMA; Edge computing; Temporal arbitration

I. Introduction

Statistics show that traffic accidents result in a staggering number of casualties annually, the majority of which are attributed to inadequate environmental perception or delayed response times. Especially in complex traffic scenarios, accurately predicting the sudden intention of the vehicle ahead (such as sudden braking and lane change) is a key factor for the intelligent assisted Driving (ADAS) rear-end collision prevention system to improve driving safety[0].

In recent years, deep learning architectures represented by Convolutional Neural Network (CNN) and Vision Transformer (ViT) have achieved remarkable results on various standardized traffic datasets[2]. However, in the deployment of real-world autonomous driving systems, on-vehicle cameras inevitably need to confront "Long-tail Scenarios" such as extreme weather (e.g., heavy rain, thick fog) and complex lighting conditions (e.g., driving at night, glare at tunnel entrances, and illuminating vehicles with high-beam illumination). In these harsh physical environments, the robustness of visual perception models often shows a precipitous drop[3].

In the frequency domain theory of digital image processing and computer vision, the Low-frequency components of an image usually characterize the smooth color transition, global illumination distribution, and macroscopic structure of the scene. While the High-frequency components densely contain sharp edges, local geometric textures, and subtle motion boundaries. The scattering effects caused by over-exposure, halo, rain and fog are represented as a High-frequency Feature Loss at the algorithm level. This loss of high-frequency information directly leads to the decrease of the model's ability to capture the edge and subtle strobe signals of the headlights[Error! Reference source not found.]. Traditional single-frame visual detection models (such as the standard architecture YOLO or the early CNN model) naturally have an inductive bias on low-frequency global context information in the network design and weight update mechanism, and ignore the extraction of high-frequency cues to a certain extent[5].

Vehicle taillights, turn signals, and hazard lamps are the core components for vehicle intention communication. However, these light signals are physically designed to have a forced fixed strobe period with a flicker frequency between 1 Hz and 2 Hz. The existing computer vision architecture, which relies on the independent detection of a single frame, is struggles to capture the periodic temporal characteristics the periodic on-off physical characteristics of the taillight[6].

Most of the existing target detection algorithms (such as Faster - R - CNN, traditional YOLO series) in processing continuous streaming data, are essentially independent static frame by frame in the implementation of the inference. The algorithm treats each frame as a temporally isolated slice. Therefore, when a video frame captured by the on-board camera happens to be in the "off" phase of the leading vehicle's turn signal, the single-frame detector will output the prediction classification of "no turning intention" due to the absence of light features in the current frame. Then, in a few milliseconds after the next frame, when a turn into the "light" phase, light spot detector can extract the high frequency characteristics, thus the output is "turn" predictions[7]. The detection

algorithm based on single frame is often because of bad environmental factors or interframe video input jitter produced by blurring and the status of the detection misjudgment. Therefore, this algorithm is rarely used in actual driving assistance[8].

In engineering practice, the traditional solution is usually to introduce a confidence threshold-based post-processing logic or a simple Moving Average algorithm to smooth these jumps. Through a ring buffer in fixed length, frame as in the past 10 or 30 frame to arithmetic average or weighted attenuation of test results, to improve the jump phenomenon[9]. However, traditional post-processing methods struggle with varying flicker frequencies, transmission delays, and prolonged occlusions. Specifically, simple sliding windows fail to capture the underlying semantic patterns of physical strobe cycles. Consequently, these methods offer limited smoothing effects and introduce unacceptable decision delays, failing to fundamentally resolve the strobe-hopping issue[10].

Therefore, this project aims to explore a vehicle intention perception framework based on YOLO combined with C2F_EMA and LSTM temporal network. By fusing spatial feature enhancement and deep temporal modeling, it compensates for the deficiencies of single perception means and improves the recognition performance of various intentions in road scenes.

II. Core Methods

Overall system architecture

The edge-side vehicle intention perception system proposed in this paper consists of two main stages: a spatial feature extraction stage based on the enhanced object detection network, and a temporal arbitration stage based on physical priority. Firstly, the input video stream is fed into the YOLO detection network fused with the C2F_EMA module frame by frame, and the bounding box containing the vehicle's spatial coordinates and preliminary intention states (such as normal, emergency, turning, etc.) is extracted. Subsequently, caused by light temporally unstable state sequences caused by light flickering was sent into lightweight temporal smoothing. The smoother eliminates the inter-frame jump through the sliding window mechanism and the state priority logic, and finally outputs the vehicle intention that is stable and conforms to the safe driving logic, which is consumed by the downstream advanced Driver Assistance System (ADAS) decision module. The flowchart of the specific project is shown in Fig. 1.

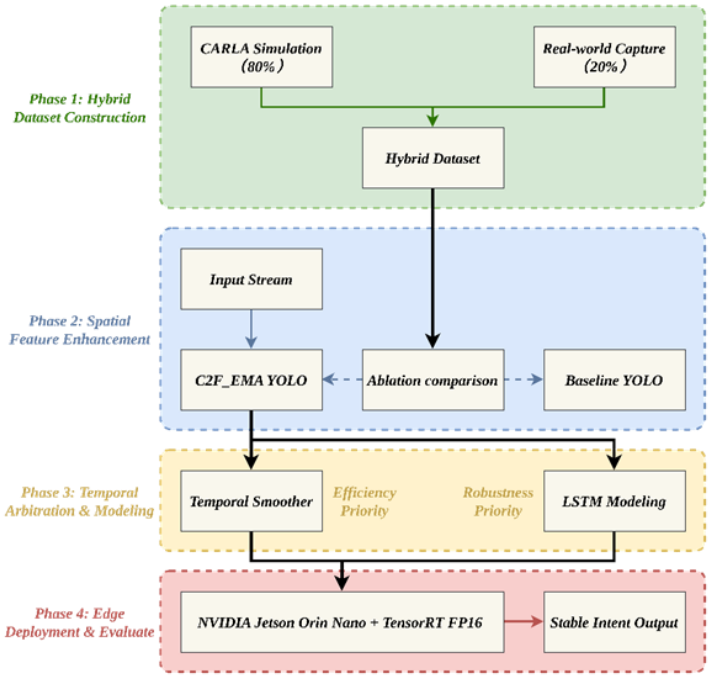

Figure 1. Architecture of vehicle intention monitoring system based on improved YOLO and LSTM

Feature extraction enhancement based on C2F_EMA

In complex nighttime long-tail scenes or bad weather, vehicle taillights are often accompanied by severe Bloom spread, overexposure, or contrast attenuation. This physical interference will cause the baseline YOLO network to lose the key high-frequency activation region in the deep feature map, which will lead to the missed detection of small targets or distant vehicles. In order to strengthen the ability of the network to represent luminescence features, a fusion module of cross-stage Local Network (C2F) and Exponential Moving Average Attention mechanism (EMA) is introduced into the Neck network of the feature pyramid. The internal logical

topology of C2F-EMA is shown in Fig. 2.

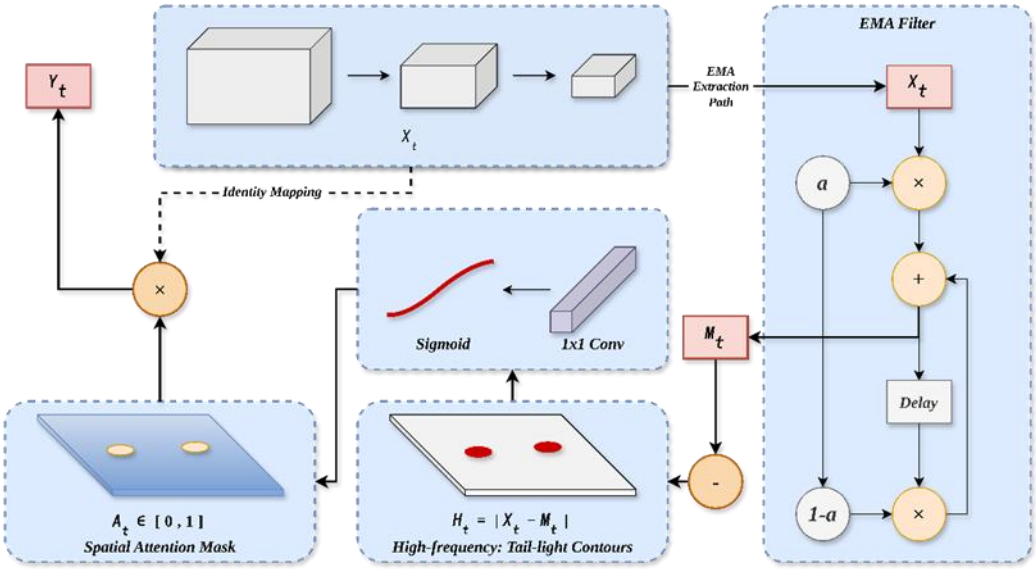

Figure 2. Internal topology of C2F_EMA module

The C2F module effectively retains the fine-grained features at multiple scales by designing rich gradient flow branches. On this basis, the EMA module is used to dynamically reshape the feature weight allocation in the channel and spatial dimensions. EMA first performs Global Average Pooling on the spatial dimensions of the input feature map to aggregate the global context information of each channel to generate a channel description, where the global context for the $c-th$ channel is computed as follows:$\boldsymbol{X} \in \mathbb{R}^{\mathbb{C}\times\mathbb{H}\times\mathbb{W}}\mathbf{z} \in \mathrm{R}^{\mathbb{C}\times 1\times 1}\mathrm{c}$

$$z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X_c}(i,j) \tag{1}$$

Subsequently, in order to capture the cross-channel interaction dependencies without significantly increasing the number of model parameters, the aggregated feature vectors are processed by the module using 1D Convolution and passed through a Sigmoid activation function σ.The attention weight vector for each channel is generated as follows.ω

$$\omega = \sigma\left(\mathrm{Conv1D}_k(\boldsymbol{z})\right) \tag{2}$$

Where, represents the kernel size of the one-dimensional convolution, which is used to control the coverage of local cross-channel interactions. k Finally, the original feature map is element-by-element multiplied along the channel dimension with the calculated dynamic weights to complete the adaptive calibration of features:$\boldsymbol{X}\omega$

$$\boldsymbol{Y} = \omega \otimes \boldsymbol{X} \tag{3}$$

Through the above reweighting mechanism, the C2F_EMA module forces the network to actively suppress large areas of irrelevant background noise (such as street lights and reflective spots), and accurately lock the receptive field in the core area of the red and yellow halo representing braking and steering, which fundamentally improves the spatial perception lower limit of the model under extreme light conditions.

At the same time, the C2F_EMA module introduces the spatial attention weight while retaining the cross-channel information interaction through the EMA structure, thereby enhancing the ability of the model to capture the subtle deformation of car lights under complex light and shadow at night

Lightweight Temporal smoother for edge deployment

The vehicle's turn signal has an inherent physical strobe cycle with the hazard warning light (double flash). This characteristic makes the single-frame object detector inevitably produce drastic jumps in detection classes between "bright" and "dark" periods when processing continuous video streams, such as repeatedly fluctuating between "left turn" and "normal" states.

In order to test the specific performance of using Smoother in edge computing devices, this paper designs a lightweight Temporal Smoother based on state statistics and Safety-Priority Arbitration. The temporal smoother maintains a memory queue with a sliding window of length, which stores the set of categories defined in the preliminary detection labels of the current time and previous frames. $N\mathcal{W}_t t\ N-1 y_i \in \mathcal{CC} =$ {Hazard,Brake,Turn_Left,Turn_Right,Normal}The memory queue is represented as follows.

$$\mathcal{W}_t = \{y_{t-N+1}, y_{t-N+2}, \cdots, y_t\} \tag{4}$$

At each time step, the system first counts the occurrence frequency of each intention category within the sliding window, defined as the sum of the indicator functions:$tS(c)\mathbb{I}$

$$S(c) = \sum_{i=t-N+1}^{t} \mathbb{I}(y_i = c), \quad \forall c \in \mathcal{C} \tag{5}$$

In order to ensure driving safety, the system introduces the anti-collision safety priority mapping mechanism, and establishes the absolute arbitration criteria for dealing with emergency situations: danger alarm (double flash) emergency braking (emergency brake) steering intention to drive normally. $P(c) >>>$The arbitration module checks whether the cumulative frequency exceeds the preset anti-interference confidence threshold according to the order of decreasing priority $S(c)\tau_c$

The stable intention extraction logic for the final output is as follows:$\widehat{Y}_t$

$$\widehat{Y}_t = \arg \max_{c \in \mathcal{C}_{ordered}} \{P(c) \mid S(c) \geq \tau_c\} \tag{6}$$

The timing arbiter filters the "dark period" jump caused by stroboscopic with very low time complexity of. $\mathcal{O}(N)$While maintaining the high accuracy of intent recognition, the computing power pressure of edge devices is reduced to a certain extent, so that the end-to-end perception system runs in a low-latency state.

LSTM-based Deep Temporal Intention Modeling Network

While the smoother addresses explicit class jumps, it essentially 'discards information' (filtering), whereas the LSTM 'exploits information' (memorizing). Moreover, although lightweight temporal smoothers can filter high-frequency information, their nature is based on hard logical statistics of surface labels, which lacks a deep understanding of the evolution of vehicle temporal characteristics, such as the gradual change process of brake lights turning on and off. To this end, this paper further introduces Long Short-Term Memory (LSTM) network to deeply model the temporal features of the target sequence.

By introducing Cell State and a sophisticated Gate Mechanism, LSTM can effectively memorize the "bright period" characteristics of car lights strobe. Given the target feature sequence output by the YOLO network at a given time, the core update process of the LSTM unit is as follows: $t\ x_t$

1. Forget Gate: Determines how much of the cell state from the previous moment needs to be preserved.C_{t-1}

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \tag{7}$$

2.Input Gate and candidate memory: Determines how much of the current input needs to be updated into the cell state.x_t

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{9}$$

3. Cell State Update: Combines the forget gate with the input gate to update the long-term memory of the current moment. C_tEven in the "dark period" (weak intent feature) when the lights go out, the network can still keep the "turn" or "rush" memory of the previous moment.$x_t f_t$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \tag{10}$$

4. Output Gate and hidden state: The hidden state of the final output is determined based on the current cell state for the final intention classification.h_t

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{11}$$

$$h_t = o_t * \tanh(C_t)$$

Through the above gating operation, the LSTM network gives the system a strong temporal tolerance to extreme stroboscopic and partial occlusion, and solves the problem of intention break between frames from the feature dimension to a certain extent.

Fig. 3 shows the logical flow comparison between the Temporal Smoother and the traditional LSTM.

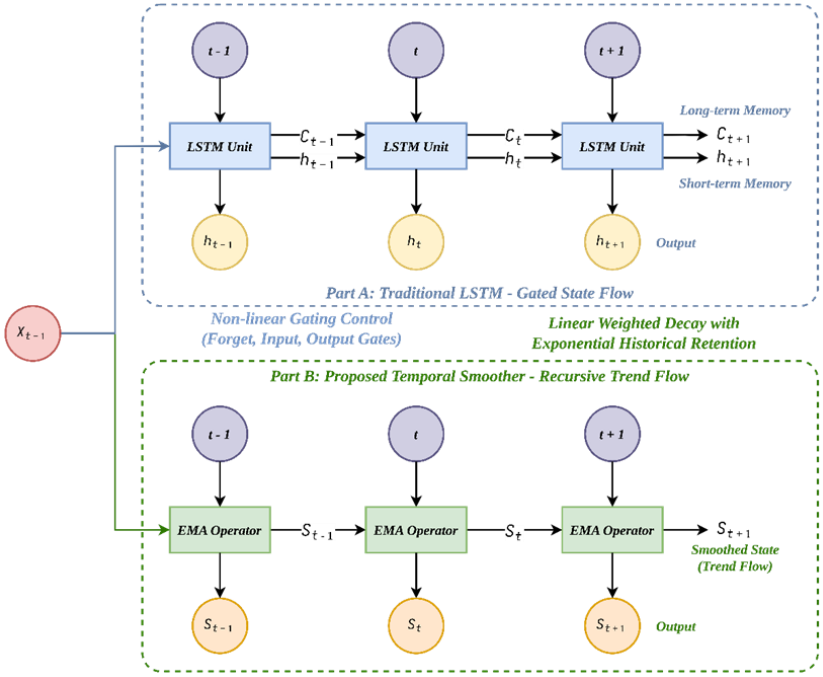

Figure 3. Logical flow comparison of Temporal Smoother and traditional LSTM

III. Experimental Setup And Dataset

Construction of mixed data set combining virtuality and reality

In the development of intelligent driver assistance systems, there is a lack of data on long-tail scenes (such as polar nights, sudden brakes and strobe in heavy rain). To this end, this paper proposes a hybrid dataset construction method based on 3D physical projection and semi-supervised real vehicle data mining.

Firstly, in the generation stage of Source Domain data, we introduce a multi-dimensional Domain Randomization strategy based on CARLA, an autonomous driving simulation engine. By dynamically injecting extreme meteorological parameters (e.g., cloud cover, precipitation, solar elevation Angle) and extreme optical parameters (e.g., ultra-high halo intensity Bloom, large aperture out of focus) into the rendering layer, the visual priors of the model for clear day environments are forced to be broken. In order to eliminate the quadratic spatial error caused by traditional manual annotation, we discard the conventional pseudo-label mechanism and directly extract the 3D Bounding Box world coordinates of the target vehicle in the underlying physics engine. Based on the pinhole imaging model of the camera, the 3D homogeneous coordinates are extracted $\boldsymbol{P_w} = [X_w, Y_w, Z_w, 1]^T$

Accurately projected to 2D pixel coordinate system, the mathematical transformation formula is as follows:$\boldsymbol{p_{img}} = [u, v, 1]^T$

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K[R \quad t]\begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{12}$$

Where is the depth scaling factor of the target; $s[\boldsymbol{R} \mid \boldsymbol{t}]$Is the Extrinsic Matrix of the camera, which represents the rotation and translation transformation from the world coordinate system to the camera coordinate system; $\boldsymbol{K}$For the camera's internal Matrix (the Intrinsic Matrix), by the focal length of the camera () () and optical center:$f_x, f_y c_x, c_y$

$$\mathbf{K} = \begin{bmatrix} \mathrm{f_x} & 0 & \mathrm{c_x} \\ 0 & \mathrm{f_y} & \mathrm{c_y} \\ 0 & 0 & 1 \end{bmatrix} \tag{13}$$

Through the above homogeneous coordinate transformation, the system realizes the automatic label generation with low error. Secondly, in order to avoid the recognition difference between pure virtual data and the real world, this paper introduces a certain amount of real traffic recorder data in the target domain. The baseline large model pre-trained on a large-scale general data set is used to extract frames with high confidence from real car videos, and the labels are manually fine-tuned, so that the real sensor noise and background interference features are extracted by the model. Finally, the simulation data and real vehicle data according to speak of mixed, build the intention covers five categories (normal, brakes, turn left, turn right, double flash) verification of the set. The diagram of label framing is shown in Fig. 4. Fig. 4 (a) shows the labeling diagram of CARLA, and the Fig. 4 (b) is the labeling diagram of actual vehicle.

Figure 4 (a). Schematic diagram of CARLA annotation

Figure 4. (b) Schematic annotation of the actual picture

Experimental environment and implementation details

The experiments in this paper are divided into two phases: cloud model training and edge-side real vehicle deployment. In the model training phase, NVIDIA Tesla P100 GPU (16GB VRAM) is used as the hardware platform. The network was built based on the deep learning framework, and the uniform resolution of the input image was set as. 640×640In terms of hyperparameter configuration, the Batch Size of model training is set to 32, the total number of training Epochs is set to 60, and the Early Stopping mechanism (Patience=20) is introduced to prevent overfitting. The initial learning rate is set to 0.01. The augmentation model is used to improve the robustness of the model to complex spatial locations and illumination distortions. Mosaic Mosaic augmentation and high-frequency perturbation of HSV color space are enabled during training. The specific parameter table is shown in Tab. 1:

Table 1. Training parameters for YOLOv8 taillight recognition

Parameter Name	Numeric
Epoch	60
Patience	20
Batch	32
Imgsz	640

In the edge deployment stage, in order to verify the real-time performance of the system in the real vehicular environment, the trained PyTorch weights are exported and deployed to the NVIDIA Jetson Orin Nano edge computing platform with limited computing power. By introducing the TensorRT inference acceleration engine, the full precision (FP32) network is quantization compiled for half precision (FP16), which maximizes the use of its computing power without loss of detection accuracy.

Evaluation metrics

In order to comprehensively evaluate the comprehensive performance of the perception system in terms of algorithm accuracy and engineering implementation, this paper uses the Mean Average Precision (mAP) and the inference efficiency index suitable for edge computing.

In terms of precision, the precision (P) and recall of a single class under a specific Intersection over Union (IoU) threshold are first calculated, and then the average precision of the class is obtained by integrating the P-R curve. The final model overall precision mAP is the arithmetic average over all classification subsets. In this paper, we focus on mAP@0.5, which is the average precision at IoU threshold of 0.5, as the core detection index. In terms of engineering efficiency, we focus on measuring the end-to-end inference frame Per Second (FPS) and single frame processing Latency of the system on the Jetson Orin Nano platform. The latency includes the whole process time from video stream image decoding, C2F_EMA network forward propagation, NMS post-processing, and finally the stable intention output by the timing smoother, which is used as a standard to measure the real-time performance of the ADAS avoidance system.

IV. Experimental Results And Discussion

Detection performance analysis and network ablation experiments

In order to verify the substantial improvement of the feature extraction ability of the baseline object detection network by the proposed C2F_EMA module, we conduct rigorous ablation experiments on the constructed hybrid validation set. The experiments focus on comparing the average precision (mAP@0.5) of the baseline YOLO model and the enhanced model after introducing C2F_EMA on various intonations (Normal, Brake, Turn_Left, Turn_Right, Hazard). The specific parameters are shown in Tab. 2 and Tab. 3

Table 2. Table of values in extreme scenarios

Model	Normal	Brake	Left	Right	Hazard	mAp50
v8	0.662	0.654	0.583	0.579	0.568	0.609
v8+C2F_EMA	0.684	0.703	0.621	0.618	0.607	0.647
Improvement	+3.3%	+7.5%	+6.5%	+6.7%	+6.8%	+6.2%

Table 3. Table of values in better light

Model	Normal	Brake	Left	Right	Hazard	mAp50
v8 (Clear)	0.754	0.715	0.742	0.708	0.721	0.718
v8+C2F_EMA (clear)	0.784	0.743	0.771	0.738	0.767	0.747
Improvement	+4.0%	+3.9%	+3.9%	+4.2%	+6.4%	+4.0%

Experimental results show that both the baseline model and the enhanced model show high detection accuracy in simple scenes with daytime or good lighting conditions. However, when the extreme long-tail samples (such as strong halo at night, small car lights at a distance, and reflection on rainy days) are tested in the hybrid validation set, the performance of the baseline model decreased significantly, especially in the sensitive category of "Brake", the miss rate increased significantly.

In contrast, the network with the introduction of C2F_EMA module achieves a certain improvement in the overall mAP@0.5. Because the Exponential Moving Average (EMA) attention mechanism dynamically reshaped the channel weights, the model successfully suppressed the activation of large area background noise such as street lights and water reflection, and accurately focused the effective receptive field of the network on the red and yellow luminous pixel area. Due to the sharp brightness and contrast mutation when the Brake light is turned on, the attention mechanism of C2F_EMA can more keenly capture this nonlinear feature gain, so the mAP increase in the Brake category is the most significant (+7.5%). The data show that the enhanced models are significantly improved, which reflects the effectiveness of the spatial feature enhancement strategy in overcoming extreme illumination distortion to a certain extent.

Performance comparison of edge-end deployment and timing module

In practical Driver assistance systems (ADAS), the delay information of the algorithm still needs to be considered while measuring the accuracy of the algorithm. In order to verify the engineering value of the timing processing module on the vehicular edge device, this paper conducts an end-to-end performance benchmark on the NVIDIA Jetson Orin Nano platform. The inference efficiency (FPS) and end-to-end Latency (Latency) of C2F_EMA based lightweight time series smoother and C2F_EMA based LSTM time series network were compared. The specific effect diagram of the Smoother logic judgment is shown in Fig. 5, the effect diagram of the LSTM logic judgment is shown in Fig. 6, and the recognition link delay is shown in Tab. 4

Figure 5. Effect of Smoother logic judgment

Figure 6. Effect diagram of LSTM logic judgment

Table 4. Link delay and accuracy of edge computing devices

Parameter Name	Value
Video Stream (Smoother) Video Stream (LSTM)	26.8ms 34.2 ms
Local Video (Smoother)	12.6ms
Local Video (LSTM)	15.6ms
AP (Smoother)	53.7%
AP (LSTM)	80.2%

The test results show that although the introduction of LSTM network can effectively extract the timing features of long sequences and solve the intention jump between frames, it will increase the computing power overhead of Jetson platform to a certain extent because it needs to read, write and update the Hidden States of the video memory when processing the sequences. Under this architecture, the single-frame latency of LSTM is increased from 26.8ms to 34.2ms (an increase of about 7.4ms), but the robustness of the higher-dimensional image is exchanged, which represents a justifiable trade-off between computational overhead and inferential robustness in ADAS applications. However, thanks to the quantization acceleration and operator optimization of TensorRT FP16 and the video input method of GStreamer, the full-link sensing system equipped with LSTM module is still able to maintain the end-to-end frame rate at about 30 FPS, and the increase in single frame delay is completely at an acceptable level.

In contrast, the temporal smoother, leveraging a deterministic heuristic based on sliding windows, exhibits lower computational complexity and state statistics, and shows certain deployment advantages in terms of low latency. However, the temporal smoother essentially relies on the hard logic arbitration of surface labels. When dealing with complex working conditions, such as irregular strobe of car lights and gradual change caused by partial occlusion, it lacks the processing of deep evolution of temporal features, and has a certain lack of recognition accuracy. As shown in the figure, under the simple recognition condition of the left picture in Fig. 5, because the lights of the front vehicle are dark for a short period of time, the timing smoother outputs that the vehicle is in a normal state according to the established logic. As shown on the right of the figure, when the model recognizes the rear vehicle due to the jitter of the transmission stream in the fast-moving video stream, the timing smoother outputs the brake indication based on the simple detection information, while the model using LSTM outputs the vehicle in the normal straight state due to the learning of the frames before and after.

In summary, although the use of LSTM module will reduce the highest detection frame rate to a certain extent, it achieves a higher dimension of intention coherence and perceptual robustness by virtue of its long short-term memory ability. This compromise of computing power within a controllable range makes up for some hard logic defects of the smoother, and proves that deploying LSTM at the edge end for deep temporal modeling has certain engineering desirability and core value. At the same time, considering that the reaction time of human drivers is about 200ms-500ms, the full-link delay of the system of 34.2ms reserves a sufficient safety margin, which proves the practical engineering significance of the scheme in ADAS avoidance scenarios.

V. Conclusion

In this paper, we propose an end-to-end perception framework combining improved YOLO and deep temporal modeling to solve the problem of insufficient robustness of intention perception of autonomous vehicles and high vulnerability to headlight strobe interference in complex long-tail scenes. In the spatial feature extraction stage, the C2F_EMA module is introduced into the network to dynamically reshape the feature weights, effectively suppress the background noise, and significantly improve the target capture ability and single frame detection accuracy of the model under extreme light conditions such as night halo and bad weather.

At the temporal logic processing level, aiming at the intention jump between frames caused by car lights strobe, this paper compares the practical engineering performance of the temporal smoother based on hard logic and Long Short-Term Memory network (LSTM). It is confirmed that although the heuristic temporal smoother has an extremely low time complexity, it is prone to intention reset in the face of complex occlusion and irregular strobes, resulting in a decrease in recognition rate. In contrast, the introduction of LSTM network for deep temporal modeling can extract and maintain the feature evolution process for a long time. The real vehicle verification on the Jetson Orin Nano edge computing platform shows that although LSTM increases a certain amount of computing power overhead, it can still maintain the end-to-end reasoning within a reasonable response time, and achieve much higher intention recognition rate and decision coherence than the smoother under most conditions.

In summary, the proposed perception scheme greatly improves the accuracy and robustness of intention recognition within an acceptable computational delay, and addresses the limitations of traditional single-frame vision and simple post-processing logic. Combined with the virtual-real mixed data set constructed in this paper, the end-to-end system provides a solution with high engineering value for the safety avoidance of Advanced Driver Assistance Systems (ADAS) on real open roads.

Acknowledgments

The authors gratefully acknowledge the financial supports by the Cooperative Scientific Research Project of the Chunhui Program, Ministry of Education, China (grant No. 202201602), Changzhou Science and Technology Project (grant No. CQ20254009, No. CJ20252018)

References

[1] Plainis S, Murray I J, Pallikaris I G. Road Traffic Casualties: Understanding The Night-Time Death Toll[J]. Inj Prev, 2006, 12(2): 125-128. Doi:10.1136/Ip.2005.011056.

[2] Pettersen F, Zhu H. Robustness Of Object Detection Of Autonomous Vehicles In Adverse Weather Conditions[Eb/Ol]. (2026-02)[2026-03-19]. Https://Doi.Org/10.48550/Arxiv.2602.12902.

[3] Chen Z, Zhang Z, Su Q, Et Al. Object Detection For Autonomous Vehicles Under Adverse Weather Conditions[J]. Expert Systems With Applications, 2025, 296: 128994. Doi:10.1016/J.Eswa.2025.128994.

[4] Gakhar I, Gupta A, Guha A, Et Al. Fourier Domain Adaptation For Traffic Light Detection In Adverse Weather[C]// Proceedings Of The Ieee/Cvf International Conference On Computer Vision. 2025: 816-825.

[5] Wang Y, Fu T, Zhou Y, Et Al. Tenet: Attention-Frequency Edge-Enhanced 3d Texture Enhancement Network[J]. Sensors, 2025, 25: 715. Doi:10.3390/S25030715.

[6] Frossard D, Kee E, Urtasun R. Deepsignals: Predicting Intent Of Drivers Through Visual Signals[C]// 2019 International Conference On Robotics And Automation (Icra). Ieee, 2019.

[7] Zhu H, Wei H, Li B, Et Al. A Review Of Video Object Detection: Datasets, Metrics And Methods[J]. Appl Sci, 2020, 10: 7834. Doi:10.3390/App10217834.

[8] Kang K, Li H, Yan J, Et Al. T-Cnn: Tubelets With Convolutional Neural Networks For Object Detection From Videos[J]. Ieee Transactions On Circuits And Systems For Video Technology, 2016. Doi:10.1109/Tcsvt.2017.2736553.

[9] Chen K, Wu G. The Vehicle Intention Recognition With Vehicle-Following Scene Based On Probabilistic Neural Networks[J]. Vehicles, 2023, 5(1): 332-343.

[10] Liu P, Qu T, Gao H, Et Al. Driving Intention Recognition Of Surrounding Vehicles Based On A Time-Sequenced Weights Hidden Markov Model For Autonomous Driving[J]. Sensors, 2023, 23: 8761. Doi:10.3390/S23218761.