vehicles

MDPI

Article

Driver Attention Region Prediction Based on Multi-Attention Mechanism Multi-Scale Fusion Network

Yunxing Chen [1,2], Guo Yu [1,2,*], Kunhui Li [1,2] and Xingyu Yuan [1,2]

[1] Hubei Key Laboratory of Vehicle-Infrastructure Collaboration and Traffic Control, Hubei University of Arts and Science, Xiangyang 441053, China
[2] School of Automobile and Traffic Engineering, Hubei University of Arts and Science, Xiangyang 441053, China
* Correspondence: y872399655@163.com

Abstract

In driver attention zone prediction tasks, accurately identifying and locating the driver's attention zone is crucial. Traditional models have significant limitations in complex driving scenarios due to their failure to fully utilize multidimensional driving environment information. To address these issues, this paper proposes a multi-attention feature fusion network (MAFF-HRNet) for driver attention region prediction. The proposed network combines high-resolution feature extraction with bimodal RGB–semantic inputs, multi-scale feature fusion, attention-based feature refinement, and temporal modeling. The experimental results on the DR(eye)VE dataset show that MAFF-HRNet improves driver attention region prediction under the current evaluation protocol. These results indicate that semantic scene information, multi-scale spatial representation, and temporal context are beneficial for generating more accurate driver attention heatmaps in complex driving scenes.

Keywords: driver attention; multi-scale; multi-attention

1. Introduction

Traditional driver attention region prediction (DARP) methods often rely on single-modal information, such as RGB images, eye-tracking data, or vehicle sensor signals. These methods can achieve certain results in relatively simple or static driving scenarios, but due to the limited dimensionality of the input information, they are unable to fully characterize the complex semantic structures and dynamic changes in real traffic environments. Therefore, their prediction performance is often significantly limited in complex driving scenarios [1]. With the development of deep learning technology, researchers are gradually trying to introduce richer visual representations and learning models to improve the ability to model the distribution of driver attention.

Although multimodal methods have made some progress in the DARP task, existing research still has shortcomings in feature modeling. On the one hand, some methods ignore the importance of features at different scales when introducing deep network structures and fail to effectively integrate the scale differences between distant and near targets. For example, [2] extracts features based on CNNs and performs simple multimodal fusion but does not systematically model multi-scale information, which is prone to misjudgment in complex traffic scenarios. On the other hand, even if some methods [3] introduce multi-modal input, they often fail to fully model the dependencies between features at different resolutions, resulting in the unstable performance of the model in dynamic scenarios.

Check for updates

Academic Editor: Lie Yang

Received: 8 May 2026
Revised: 27 June 2026
Accepted: 3 July 2026
Published: 5 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

In addition, driver attention prediction is essentially a time-dependent problem, as the driver's attention area changes continuously over time. Although some studies [4] have attempted to model temporal information using CNN-LSTM structures, in complex driving environments, these methods often fail to simultaneously take into account cross-scale spatial features and temporal dependencies. The lack of scale information or insufficient temporal modeling limits the model's ability to characterize the true distribution of driver attention in high-speed or complex traffic scenarios.

In summary, the DARP task still faces many challenges in complex traffic environments. Considering the dynamic characteristics of the driving environment and the task-driven nature of driver attention, further improving DARP performance requires systematic research from multiple aspects, including feature representation methods, multi-scale information fusion, and temporal modeling strategies. Specifically, existing research mainly focuses on the following three key issues: (1) In complex driving scenarios, image features of different scales and resolutions carry key information at different levels. High-resolution features help capture fine-grained targets such as distant pedestrians and traffic signs, while low-resolution features are more conducive to modeling the overall road structure and scene layout. However, many existing methods rely on a single resolution during feature extraction or obtain high-level semantic features through multiple downsampling, resulting in the loss of spatial detail information, thus affecting the accurate localization of the driver's attention region. This problem is particularly prominent in high-speed driving or complex traffic environments, making it difficult for the model to simultaneously consider local details and global structural information. (2) The formation of the driver's attention region is influenced by multiple factors, including semantic information, spatial structure, and dynamic behavioral characteristics that change over time. The driver's attention region is not only determined by the visual input at the current moment but also closely related to traffic conditions and driving behavior in historical frames. Furthermore, in real-world driving, the driver's attention often dynamically switches between distant and near targets; relying solely on information from a single scale or isolated frames is insufficient to accurately depict this process. Therefore, effectively integrating multi-scale spatial features with temporal contextual information within a unified framework is one of the key issues for improving the stability and accuracy of DARPs. (3) RGB–semantic bimodal DARP models usually introduce additional training difficulty because of the distribution gap between RGB images and semantic maps. Training from random initialization may lead to slow convergence and optimization instability. Therefore, how to use prior knowledge or pre-training strategies to improve optimization stability and prediction performance is an important issue in DARP research. It should be noted that this study focuses on prediction accuracy and training stability, rather than real-time deployment.

This paper proposes a unified multi-scale, RGB–semantic bimodal, and temporal modeling framework based on the task requirements of driver attention region prediction. This framework, grounded in high-resolution feature representations, aims to simultaneously consider fine-grained spatial information, global semantic structure, and temporal continuity in complex and dynamic driving scenarios. Specific contributions are as follows:

(1) We construct an RGB–semantic bimodal fusion framework for driver attention region prediction based on HRNet. Instead of using a single low-resolution representation, the framework preserves high-resolution spatial details and integrates multi-scale contextual information, which is useful for localizing attention regions associated with small, distant, or structurally important traffic objects.
(2) We introduce a cross-modal and spatial attention refinement process to improve the interaction between RGB features and semantic features. CMA is used to model the relationship between visual appearance and semantic scene information, while DSA

further enhances spatially important regions. This design helps the model focus on traffic-related areas that are more relevant to driver attention.

(3) We incorporate ConvLSTM and a prior-knowledge-based training strategy into the RGB–semantic bimodal framework. ConvLSTM is used to model short-term temporal consistency between consecutive frames, while RGB-only pre-training provides a task-related initialization for subsequent bimodal fine-tuning. This improves the optimization stability of the proposed model under the current DARP setting.

2. Related Works

In early studies, driver attention prediction was mainly inspired by the human visual attention mechanism and modeled using bottom-up saliency detection methods. For example, the classic saliency model proposed by Itti et al. [5] generates a saliency map by calculating the contrast relationship between low-level features such as color, brightness and direction. However, such methods have significant limitations in driving scenarios. On the one hand, driver attention has obvious task-driven characteristics, and its distribution is often closely related to traffic rules, risk avoidance and driving goals, rather than being determined solely by visual saliency [6]. On the other hand, these methods are usually based on static image modeling, which makes it difficult to characterize the attention shift process that changes over time in dynamic driving scenarios. With the development of computer vision and large-scale labeled data, researchers have gradually turned to deep learning-based methods to improve the model's expressive power for complex scenarios [7].

Deep learning-based DARP methods typically utilize convolutional neural networks (CNNs) to automatically extract visual features from RGB images and combine them with recurrent neural networks (such as LSTMs) to model the temporal changes in driver attention. A typical example is DR(eye)VE [8], which constructs a CNN–LSTM structure to predict driver attention distribution by fusing road scene images with eye-tracking data. Although this type of method introduces a temporal modeling mechanism to some extent, its ability to model complex traffic semantic information is still limited because it mainly relies on a single modal input. Similarly, some studies based on physiological signals or vehicle sensor data can capture changes in driver state, but they often lack a comprehensive understanding of multidimensional information about the driving environment [9]. Therefore, methods that rely solely on a single data source are prone to problems such as unstable prediction or a shift in the area of interest in high-speed driving or dense traffic scenarios.

To overcome the shortcomings of single-modal methods, recent studies have gradually introduced multimodal information, which enhances the model's understanding of driving scenarios by fusing visual data with vehicle state information, semantic information, or other auxiliary modalities. For example, Chen et al. [10] proposed FBLNet, which combines visual information and vehicle state data and improves the accuracy of driver attention prediction through a feedback mechanism; Hu et al. [11] used unlabeled eye-tracking data and scene features to perform multimodal fusion, achieving more robust attention estimation; in addition, some studies [12] have also explored unsupervised or weakly supervised multimodal fusion methods to reduce dependence on labeled data and improve model efficiency. These methods show that multimodal information has a positive effect on improving DARP performance.

In the further development of multimodal methods, some studies have introduced more complex network structures and attention mechanisms to enhance information interaction between different modalities. For example, [13] combined convolutional neural networks with Transformer structures and modeled the relationship between multimodal features through self-attention mechanisms, thereby improving the accuracy of driver attention region prediction. In addition, some works [14] have introduced RGB images,

semantic segmentation maps, depth maps and optical flow information at the same time to comprehensively characterize driving scenarios. However, the introduction of multimodal inputs often significantly increases the model training cost and computational complexity. To this end, some studies [15] have tried to use transfer learning or pre-training strategies to improve training efficiency.

To address the above issues, existing research has explored various approaches. In terms of multi-scale feature representation, HRNet proposed by Zhou et al. [16] has achieved good results in complex image segmentation tasks by maintaining multi-resolution feature representation in parallel. Wang et al. [17] introduced HRNet into the object detection task and verified its ability to effectively preserve fine-grained features under multi-scale conditions, especially in complex environments with dense pedestrian and vehicle traffic. Xu et al. [18] further applied HRNet to the dynamic driver attention prediction task, showing that it can better capture the dynamic changes in the driver's attention area in high-speed driving scenarios. Meanwhile, some studies have enhanced the ability to fuse multi-scale information by introducing a feature pyramid network (FPN) and pyramid pooling module (PPM). Wang et al. [19] introduced an FPN into the driver attention region prediction model, which effectively improved the model's ability to recognize multi-scale targets. Li et al. [20] pointed out that a PPM can enhance the model's global perception ability, thereby significantly improving target recognition and localization performance in complex scenarios. Yang et al. [21] further analyzed the advantages of FPNs in integrating semantic and detail features at different levels, while Li et al. [22] achieved a more refined fusion of multi-scale features by introducing an attention mechanism into the FPN and PPM structures. However, these methods are mostly focused on static images or general vision tasks, and the modeling of temporal dependencies in driver attention prediction is still relatively limited.

In terms of temporal modeling, existing research has shown that capturing the changing patterns of driver attention between consecutive frames is of great significance for improving prediction stability. The MEDIRL model proposed by Baee et al. [23] emphasizes the importance of temporal information modeling in dynamic driving environments. Zhang et al. [24] constructed a driver attention prediction model based on LSTM, which can use historical gaze behavior to predict future attention areas. Yang et al. [25] further proposed a CNN-LSTM structure, which improved prediction accuracy in complex traffic scenarios by modeling the dependencies between multiple frames. Lee et al. [26] used a temporal convolutional network to model multiple frames of images, verifying the important role of historical information in dynamic scenes. However, while these methods enhance the ability of temporal modeling, they often fail to fully combine the expression of multi-scale spatial features.

In terms of model training strategies, some studies have introduced pre-training or transfer learning methods based on prior knowledge to improve training efficiency and prediction performance. Fu et al. [27] proposed a staged training strategy consisting of first pre-training on RGB images and then fine-tuning multimodal models, which effectively improved the prediction accuracy of DARP. Li et al. [28] used RGB image pre-trained weights to guide the initialization of multimodal models, which accelerated the convergence speed of the models. Zhao et al. [29] constructed a joint training framework by sharing prior knowledge among multimodal models, which significantly shortened the training time and improved the generalization ability of the models. In addition, Wang et al. [30] and Kim et al. [31] verified the effectiveness of prior knowledge in complex driving scenarios from the perspectives of self-supervised learning and model initialization, respectively.

3. Methods

The architecture of the proposed method is shown in Figure 1. MAFF-HRNet is designed according to the characteristics of driver attention region prediction in complex driving scenes, where the model needs to capture fine-grained spatial details, task-related semantic cues, multi-scale traffic objects, and short-term temporal changes. Therefore, each module in MAFF-HRNet is introduced with a specific functional purpose. HRNet is used to preserve high-resolution spatial details, CMA is used to strengthen the interaction between RGB features and semantic features, DSA is used to refine spatially important regions, PPM-FPN is used to integrate multi-scale contextual information, and ConvLSTM is used to model temporal consistency in consecutive frames. These modules jointly support driver attention region prediction from the perspectives of spatial detail preservation, RGB–semantic feature fusion, multi-scale feature interaction, spatial attention refinement, and temporal modeling.

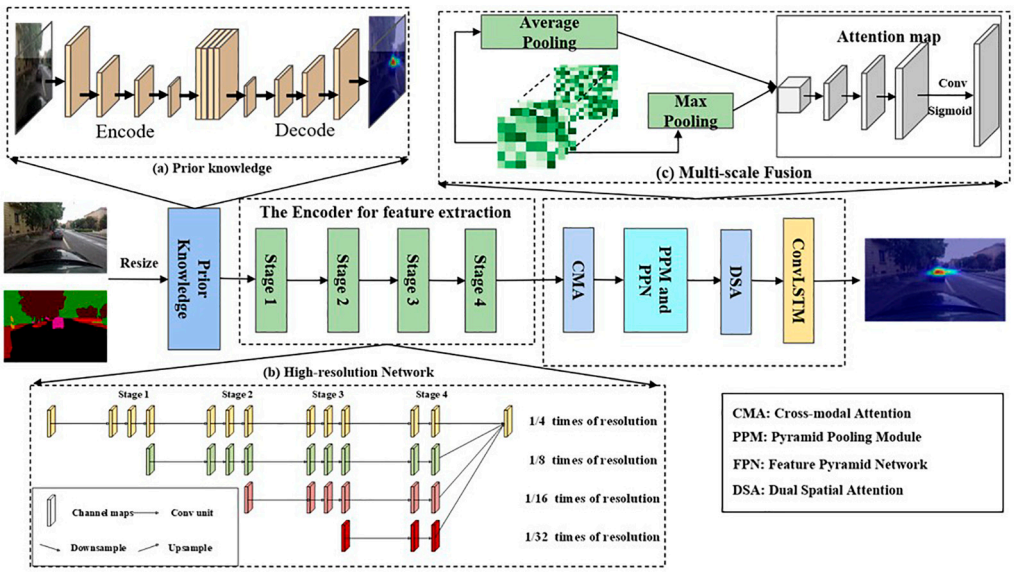


Figure 1. MAFF-HRNet architecture for DARP. Given a video segment with RGB frames, a semantic image is first obtained through a segmentation model. Then, the RGB image and semantic image are fed into HRNet for feature extraction, generating RGB and semantic features at different resolutions. Multi-scale fusion techniques combining CMA and DSA are used to fuse the features, and the final features are generated through ConvLSTM. Finally, an attention heatmap is obtained through decoding.

Given a sequence of five consecutive frames, the RGB input is resized to 192×320 and represented as $B \times 5 \times 192 \times 320 \times 3$, while the corresponding semantic input is represented as $B \times 5 \times 192 \times 320 \times 1$. The RGB and semantic sequences are first fed into the HRNet-based feature extraction branches to obtain multi-scale spatial representations. The extracted RGB and semantic features are then aligned in spatial resolution and fused by CMA to enhance cross-modal interaction. DSA is subsequently applied to emphasize spatially important regions related to driver attention. The refined features are further processed by PPM-FPN to aggregate multi-scale contextual information. ConvLSTM is then used to model short-term temporal dependencies over the five-frame sequence, and prediction is supervised using the attention map of the last frame. Finally, the decoder progressively upsamples the fused feature representation through convolution, batch normalization, nonlinear activation, and bilinear upsampling, followed by a 1×1 convolution with sigmoid activation to generate the final attention heatmap with a size of $B \times 192 \times 320 \times 1$.

3.1. High-Resolution Network Module (HRNet)

We adopted a High-Resolution Network (HRNet) module to fully leverage its advantages in high-resolution image processing, especially in the face of complex backgrounds or significant changes in lighting. This helps preserve high-resolution spatial information and reduces detail loss during feature extraction. The detailed process of the HRNet module is as follows:

Initial Convolution Layer: This layer extracts initial low-level features from the original input image. This includes convolutional layers, batch normalization (BN) layers, and the ReLU activation function.

Residual Block: Our high-resolution subnetwork employs residual blocks (ResBlocks). The key feature of residual blocks is the introduction of a residual learning mechanism, where the input is directly added to the output via skip connections. This effectively alleviates the vanishing gradient problem in deep networks and facilitates the training of deeper networks. Each residual block consists of two convolutional layers, sandwiched between BN layers and the ReLU activation function.

Parallel Multi-resolution Streams: The high-resolution subnetwork processes the highest-resolution feature maps, while other parallel low-resolution streams process lower-resolution feature maps. Information between these streams is exchanged through cross connections, ensuring that the high-resolution streams utilize the contextual information in the low-resolution streams, while the low-resolution streams benefit from the fine-grained information in the high-resolution streams.

We start with a high-resolution subnetwork as the first stage, gradually adding subnetworks from high resolution to low resolution to form more stages. We connect these multi-resolution subnetworks in parallel, and throughout the process, we perform multi-scale fusion by repeatedly exchanging information across these parallel multi-resolution subnetworks, thus obtaining richer high-resolution representations. Therefore, the predicted attention heatmap is likely to be more accurate and spatially precise.

3.2. Attention Mechanism

3.2.1. Cross-Modal Attention (CMA) Mechanism

We introduce a cross-modal attention (CMA) mechanism on the same-scale feature layer output by the HRNet backbone network. This module uses RGB features as queries and semantic features as keys and values. By explicitly modeling cross-modal correlations, it guides the modeling of visual attention with semantic information.

The RGB features and semantics output by HRNet are as follows:

$$F_r^l \in R^{H_l \times W_l \times C_l}, F_s^l \in R^{H_l \times W_l \times C_l} \quad (1)$$

where H_l, W_l, C_l represent the spatial resolution and number of channels of the l-th layer. First, channel mapping is performed using 1×1 convolutions to construct the query, key, and value:

$$Q = F_r^l W_q, K = F_s^l W_k, V = F_s^l W_v \quad (2)$$

where W_q, W_k, and $W_v \in R^{C_l x d}$ denote the attention embedding dimensions, and then the cross-modal attention weights are calculated:

$$A = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \quad (3)$$

The final cross-modal fusion features are as follows:

$$\mathrm{F}_{cma} = \mathrm{AV} \tag{4}$$

And this is fused with RGB features through residual connections:

$$\mathrm{F}_{out} = \mathrm{F}_r^l + \mathrm{F}_{cma} \tag{5}$$

All attention operations are performed at the same spatial resolution to avoid additional interpolation operations. The input and output feature dimensions of the CMA module remain consistent, providing a stable feature representation for subsequent PPM-FPN multi-scale fusion.

3.2.2. Dual Spatial Attention Mechanism (DSA)

The DSA proposed in this paper achieves a progressive enhancement in the driver's attention region by introducing spatial constraints at different stages of the network. Specifically, given the input features $\mathrm{F} \in \mathrm{R}^{H \times W \times C}$, DSA first performs average pooling and max pooling along the channel dimension:

$$\mathrm{F}_{avg} = AvgPool_c(\mathrm{F}), \mathrm{F}_{max} = MaxPool_c(\mathrm{F}) \tag{6}$$

The average-pooled and max-pooled feature maps are concatenated and processed by a convolutional layer to generate a spatial attention map.

$$\mathrm{M}_s = \sigma\left(Conv_{7\times7}\left(\left[\mathrm{F}_{avg}; \mathrm{F}_{max}\right]\right)\right) \tag{7}$$

The final output is as follows:

$$\mathrm{F}_{dsa} = \mathrm{M}_s \odot \mathrm{F} \tag{8}$$

It should be noted that DSA is not used as an isolated attention module but is embedded in both the multi-scale feature fusion stage and the decoding stage, applying spatial constraints twice to the cross-modal fused features to improve the stability of attention prediction in complex traffic scenarios.

3.3. Multi-Scale Fusion Technology

In DARP tasks, accurately identifying and locating the driver's attention focus is crucial. Traditional single-scale feature extraction methods often fail to capture sufficient detail at different scales, resulting in poor model performance in complex scenes. To address this, we propose a multi-scale feature fusion technique that combines a pyramid pooling module and a feature pyramid network.

Pyramid Pooling Module (PPM)

The PPM is located at the end of the encoder and directly acts on the highest-level semantic features $\mathrm{F} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ before ConvLSTM. The specific operation includes the following steps:

(1) Pooling operations of different scales are used to extract contextual information of different scales, resulting in four sets of feature maps P_i:

$$P_s = Pool(F), s \in \{1\times1, 2\times2, 3\times3, 6\times6\} \tag{9}$$

(2) For each pooling result P_s, dimensionality reduction is performed using 1×1 convolutions, reducing the number of channels at each pooling scale to C/4.

$$P'_s = \text{Conv1} \times 1(P_s) P'_s \in \mathbb{R}^{H_s \times W_s \times \frac{C}{4}} \tag{10}$$

P_{final} is obtained by concatenating the pooling results at different scales:

$$P_{\text{final}} = \text{Concat}\left(P'_{\text{i}}\right) \tag{11}$$

After spatial and RGB–semantic feature fusion, ConvLSTM is used to model short-term temporal consistency between consecutive driving frames. In the proposed setting, the temporal input length is set to five frames with a temporal sampling stride of 1. For a target frame t, the input sequence consists of frames {t − 4, t − 3, t − 2, t − 1, t}, and the attention map of frame t is used as the supervision target. Therefore, the model only uses the current and previous frames for prediction, without accessing future frames after the target frame. Driver attention usually changes continuously over time, and adjacent frames provide useful temporal context for predicting the current attention region. By introducing ConvLSTM after feature fusion, the model can reduce frame-to-frame attention fluctuation and capture common attention transitions in driving videos, such as gradual shifts between road regions, vehicles, pedestrians, and traffic-related objects. For extremely rapid or highly irregular attention shifts, a standard ConvLSTM may still be limited by the input sequence length and recurrent structure. Such cases may require stronger temporal modeling, such as Transformer-based temporal attention or other temporal modules.

3.4. Prior Knowledge Training Methods

It is worth noting that the proposed method uses both RGB images and semantic maps, which increases training complexity compared with single-modal input. To improve optimization stability, we adopt a prior-knowledge-based training strategy, in which an RGB-based model is first trained to provide task-related initialization for subsequent RGB–semantic bimodal fine-tuning. This strategy is designed to facilitate model training and improve prediction performance.

Specifically, the main process of the prior-knowledge-based training method is as follows:

(1) Acquisition of prior knowledge: Through preliminary research, we found that using only RGB images can effectively predict the driver's attention region. Therefore, as a first step, we pre-train the model using RGB images to obtain preliminary DARP capabilities. At this point, the model obtains a rough prediction effect, and these prediction results, together with the model's initial weights, constitute our "prior knowledge."

(2) Using prior knowledge to replace layer selection: Once the prior knowledge is obtained, we apply it to some layers of the model. This prior knowledge will serve as the initial weights for some network layers, thereby accelerating the subsequent training process. It should be noted that the RGB-only pre-training stage is used to provide a task-related initialization for the model rather than to freeze the network parameters or constrain the final prediction to RGB-only features. The final performance is therefore not determined only by the initial RGB prior, because the network is further optimized using both RGB images and semantic images in the subsequent RGB–semantic bimodal fine-tuning stage. During this process, the semantic branch and the cross-modal fusion modules continue to learn task-related semantic information, and the influence of the initial RGB prior can be adjusted through bimodal optimization. Thus, the prior knowledge obtained from RGB-only training serves as an optimization starting point, while the final attention prediction is learned through RGB–semantic bimodal fusion.

4. Experimental Results

4.1. Implementation Details

All experiments were conducted on a workstation equipped with an NVIDIA RTX 4060 Ti GPU, under CUDA 11.6, Python 3.9, and TensorFlow 2.10.0. To avoid temporal leakage, the DR(eye)VE data were partitioned using a video-wise split rather than a frame-wise random split, and each video was assigned to only one subset before sequence generation.

The semantic maps were generated offline using a pre-trained BiSeNet-Xception segmentation model. In the semantic image generation stage, each RGB frame was resized to 512 × 1024 and normalized before being fed into the segmentation model. The predicted class probability map was converted into a semantic label map using argmax and then resized back to the original frame size. The generated semantic map was saved with the same frame name as the corresponding RGB frame to ensure frame-level alignment.

During MAFF-HRNet training, all RGB frames and generated semantic maps were resized to 192 × 320 before being fed into the network. For temporal modeling, each training sample consisted of five consecutive frames, and the attention map of the last frame was used as the supervision target. The model was trained using the Adam optimizer with an initial learning rate of 1e-4. The batch size was set to 4, and the model was trained for 100 epochs.

4.2. Experimental Data

Considering that this study focuses on driver attention region prediction in complex driving environments with RGB–semantic bimodal input, we selected the publicly available DR(eye)VE dataset as the primary benchmark. The DR(eye)VE dataset contains 74 video clips and a total of 555,000 frames, covering diverse driving scenes, including city centers, rural roads, highways, different weather conditions, and illumination changes. Therefore, it provides a suitable benchmark for evaluating the proposed method under a consistent experimental protocol. Some scene classifications are shown in Figure 2. It should be noted that the current experimental validation is conducted within the DR(eye)VE dataset. Since different datasets may have different annotation protocols, scene distributions, and driving domains, cross-dataset generalization is not fully evaluated in this study.



Figure 2. Data charts under different traffic environments, weather conditions, and lighting conditions.

4.3. Evaluation Indicators

To comprehensively evaluate the performance of driver attention prediction, we adopt four widely used saliency prediction metrics, including KL-div, CC, SIM, and NSS. Specifically, KL-div measures the distribution discrepancy between the predicted saliency map and the ground-truth map, where lower values indicate better performance. CC, SIM, and NSS evaluate complementary aspects of prediction quality from the perspectives of linear correlation, distribution similarity, and fixation-based response strength, respectively, where higher values indicate better performance. These metrics are used jointly to provide a more comprehensive evaluation of the predicted attention maps.

4.4. Comparison Experiment

To evaluate the performance of the proposed MAFF-HRNet, we compare it with four representative methods, HFT, MLNet, CDNN, and SCAFNet, on the DR(eye)VE dataset. The comparisons were conducted under the same DR(eye)VE benchmark settings, and the reported results are intended to evaluate the model's performance on this dataset. These baselines were selected to cover different types of attention and saliency prediction methods, including hand-crafted visual saliency, deep visual saliency prediction, driver attention prediction, and semantic-aware driver attention modeling. The quantitative results are reported using KL-div, CC, SIM, and NSS in Table 1, and the corresponding comparison curves and qualitative visualizations are shown in Figures 3 and 4.

Table 1. Comparison of test performance of different models.

Model	KL-div	CC	SIM	NSS
HFT	2.1542	0.2955	0.1787	0.8629
MLNet	0.8455	0.6321	0.3925	1.5363
CDNN	0.4945	0.8852	0.6323	0.9521
SCAFNet	0.4952	0.8547	0.7025	1.6486
Proposed	0.3141	0.9203	0.7124	4.9875

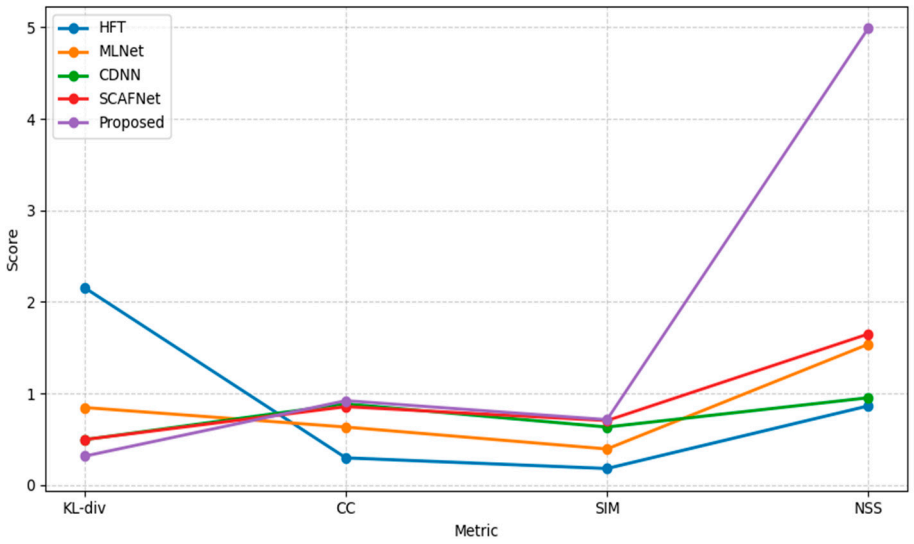


Figure 3. Comparison of line graphs for different models.

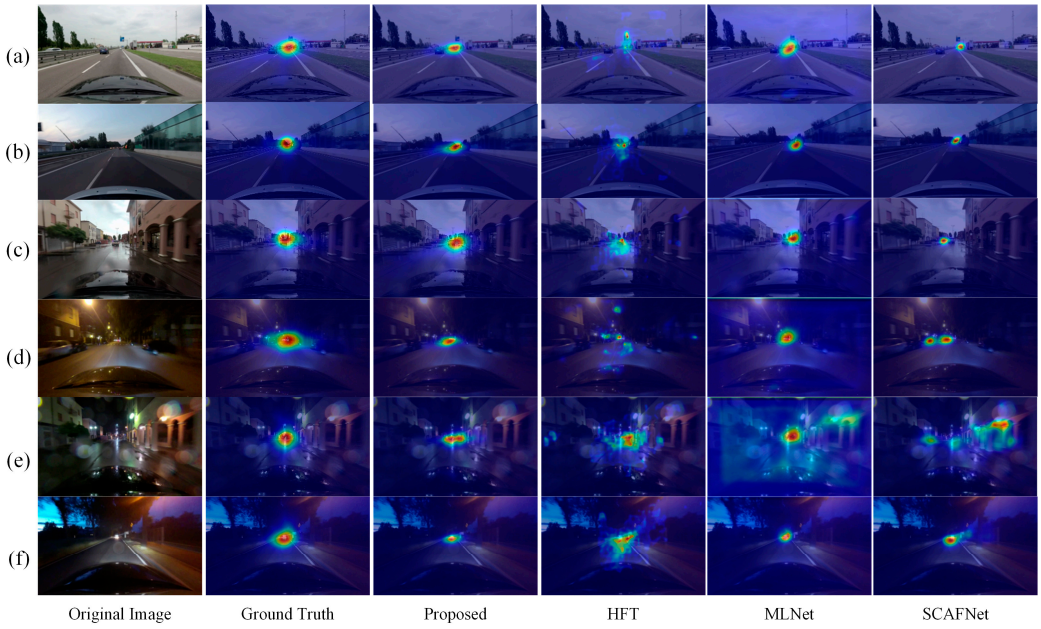


Figure 4. Visualization of different methods for DARP. Each row (**a**–**f**) shows one representative driving scene. The columns from left to right represent: original image, ground truth, prediction result of the proposed MAFF-HRNet, prediction result of HFT, prediction result of MLNet, and prediction result of SCAFNet. Warmer colors indicate higher driver-attention responses, while cooler colors indicate lower responses.

As shown in Table 1, MAFF-HRNet obtains favorable results among the compared methods under the DR(eye)VE evaluation protocol. Specifically, the proposed method obtains the lowest KL-div value (0.3141), indicating a smaller distribution discrepancy between the predicted attention map and the ground-truth gaze map. It also achieves higher CC (0.9203), SIM (0.7124), and NSS (4.9875) scores than the compared methods, indicating improved global consistency, distribution similarity, and fixation-related response. Among these metrics, NSS is a fixation-centric evaluation metric that measures the normalized response of the predicted attention map at ground-truth gaze locations and reflects the degree to which the predicted attention is concentrated around fixation-related regions. Together with the lower KL-div value and higher CC/SIM values, the NSS result provides complementary evidence that MAFF-HRNet improves driver attention region prediction under the current evaluation protocol.

As shown in Figure 4, HFT tends to generate scattered hotspots around high-contrast regions, while MLNet, CDNN, and SCAFNet can capture part of the salient regions but may still suffer from saliency drift or edge diffusion. In contrast, MAFF-HRNet produces more concentrated attention regions that are closer to the ground-truth gaze distribution in the presented examples. These improvements are mainly attributed to high-resolution feature preservation, RGB–semantic bimodal feature fusion, cross-modal and spatial attention refinement, multi-scale feature interaction through PPM-FPN, and temporal modeling with ConvLSTM. For scenes involving rapid gaze shifts, ambiguous traffic context, or imperfect semantic maps, the prediction may become less stable, and the predicted attention region may deviate from the ground-truth gaze distribution. Therefore, the results should be interpreted from the joint evidence of KL-div, CC, SIM, NSS, and qualitative visualization under the DR(eye)VE evaluation protocol.

4.5. Ablation Experiment

To investigate the contribution of different modules to DARP, we tested different combinations based on evaluation metrics. Table 2 and Figure 5 show the contribution of different

modules to DARP, with the evaluation based on reducing the use of a particular module. These include models without HRNet (No-HRNet), models without CMA (No-CMA), models without DSA (No-DSA), models without PPM and FPN (No-PPM-FPN), models without Con-vLSTM (No-ConvLSTM), and models without prior knowledge (No-PK).

Table 2. Performance comparison of different combinations.

Methods	HRNet	CMA	DSA	PPM-FPN	ConvLSTM	PK	KL-div	CC	SIM
No-HRNet	×	✓	✓	✓	✓	✓	0.3856	0.7352	0.6583
No-CMA	✓	×	✓	✓	✓	✓	0.3631	0.7453	0.6124
No-DSA	✓	✓	×	✓	✓	✓	0.3710	0.8134	0.6209
No-PPM-FPN	✓	✓	✓	×	✓	✓	0.3912	0.8347	0.6622
No-ConvLSTM	✓	✓	✓	✓	×	✓	0.4412	0.7151	0.5863
No-PK	✓	✓	✓	✓	✓	×	0.4235	0.8097	0.6219
Ours	✓	✓	✓	✓	✓	✓	0.3141	0.9203	0.7124

"✓" denotes that the corresponding module is retained in the model, while "×" denotes that the corresponding module is removed in the ablation setting.

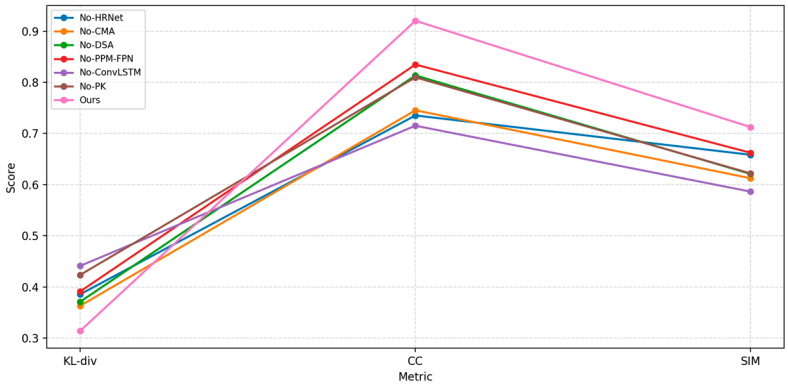


Figure 5. Ablation comparison curves of different models.

It should be noted that the current ablation study mainly focuses on module removal analysis within the proposed MAFF-HRNet framework. In the current implementation, the PPM and FPN are integrated as a coupled multi-scale fusion block, and therefore they are evaluated together as PPM-FPN rather than as two independently removed modules. In addition, although NSS is used in the main comparison experiment, the current ablation table mainly reports KL-div, CC, and SIM to analyze distribution-level prediction changes. More complete ablation studies, including NSS-based ablation evaluation and a separate analysis of PPM and FPN, will be further investigated in future work.

We explored the contribution of each component to model performance by progressively removing different modules. The experimental results show that removing any module negatively impacts model performance to varying degrees. When ConvLSTM is removed, the model fails to capture cross-frame spatiotemporal dependencies, causing KL-div to surge by over 40% and CC and SIM to drop by 22% and 18% respectively, resulting in the most significant performance degradation. This demonstrates the crucial role of temporal modeling in stabilizing attention trajectories. Removing the HRNet backbone leads to a significant drop in all metrics due to the loss of high-resolution details, highlighting the fundamental role of high-resolution features in providing fine context in complex scenes. Disabling cross-modal channel attention (CMA) weakens the complementary relationship between RGB and semantic features, causing SIM to drop to 0.61, indicating that cross-modal interaction is indispensable for suppressing semantic noise and highlighting salient regions. While removing dual spatial attention (DSA) retains a high CC, it sacrifices precise focusing on local saliency. Removing PPM-FPN weakens multi-scale contextual

aggregation, resulting in higher KL-div and lower SIM, demonstrating the importance of multi-scale information in scenes with mixed far/near targets. Finally, disabling the "RGB prior pre-training + RGB–semantic bimodal fine-tuning" strategy significantly reduces both CC and SIM, validating the value of prior knowledge in accelerating convergence and improving generalization performance.

Overall, the ablation results show that removing different components causes performance degradation to different degrees. This indicates that the introduced modules contribute to the prediction performance of MAFF-HRNet within the current framework. Therefore, architectural complexity is related to the multi-factor nature of driver attention region prediction, including spatial detail preservation, RGB–semantic feature fusion, multi-scale feature interaction, spatial attention refinement, and temporal modeling, rather than simple module stacking.

Although these results indicate the effectiveness of MAFF-HRNet for driver attention region prediction, the current model is relatively complex compared with lightweight single-frame saliency models, because it integrates high-resolution feature extraction, RGB–semantic fusion, multi-scale contextual aggregation, and ConvLSTM-based temporal modeling. Therefore, the results should mainly be interpreted from the perspective of prediction quality under the DR(eye)VE evaluation protocol, rather than as evidence of real-time in-vehicle deployment capability. The computational cost of the model should be considered when applying it to resource-constrained onboard platforms. It should be noted that the current ablation study mainly evaluates the contribution of each component through module removal experiments within the MAFF-HRNet framework. More comprehensive module replacement comparisons and an independent analysis of different input modalities would further strengthen the evaluation of module selection and semantic guidance.

5. Conclusions

In summary, this study proposes MAFF-HRNet, a bimodal RGB–semantic fusion network for driver attention region prediction. By integrating high-resolution feature extraction, cross-modal attention, dual spatial attention, multi-scale feature fusion, temporal modeling with ConvLSTM, and a prior-knowledge-based training strategy, the proposed method improves driver attention region prediction on the DR(eye)VE dataset. The results show that combining semantic scene information, multi-scale spatial representation, and temporal context is beneficial for generating more accurate attention heatmaps in complex driving scenes.

Although the DR(eye)VE dataset contains diverse driving scenarios, the current validation is still limited to this benchmark, and the generalization ability of MAFF-HRNet under unseen driving domains remains to be further examined. Future work will extend the evaluation to strictly harmonized cross-dataset settings and investigate domain adaptation strategies for adverse weather, illumination changes, and substantially different driving scenarios, such as highways and dense urban areas. Since ConvLSTM mainly captures short-term temporal consistency, more advanced temporal modules, such as Transformer-based temporal attention, will also be explored to model long-range dependencies and highly irregular attention shifts. In addition, the computational efficiency and deployment performance of MAFF-HRNet on resource-constrained onboard platforms will be further evaluated. More fine-grained ablation analysis of the multi-scale fusion components will also be considered to better understand the contribution of different architectural modules.

Author Contributions: Conceptualization, Y.C. and G.Y.; methodology, Y.C. and G.Y.; formal analysis, G.Y.; investigation, Y.C. and G.Y.; writing—original draft preparation, Y.C. and G.Y.; writing—review and editing, K.L. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by Hubei Province key Research and Development Project (2025BAB082), in part by the Hubei provincial Natural Science Foundation Joint Fund for Innovation and Development Project (2024AFD045) and in part by the Guided Project of Scientific Research Plan of Hubei Provincial Department of Education for 2024 (B2024436).

Data Availability Statement: The DR(eye)VE dataset used in this study is publicly available from the original dataset provider. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DARP	Driver Attention Region Prediction
MAFF-HRNet	Multi-Attention Feature Fusion High-Resolution Network
HRNet	High-Resolution Network
CMA	Cross-Modal Attention
DSA	Dual Spatial Attention
PPM	Pyramid Pooling Module
FPN	Feature Pyramid Network
ConvLSTM	Convolutional Long Short-Term Memory
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
KL-div	Kullback–Leibler Divergence
CC	Correlation Coefficient
SIM	Similarity
NSS	Normalized Scanpath Saliency

References

1. Chen, C.; Wang, G.; Peng, C.; Zhang, X.; Qin, H. Improved Robust Video Saliency Detection Based on Long-Term Spatial-Temporal Information. *IEEE Trans. Image Process.* **2020**, *29*, 4096–4110. [CrossRef] [PubMed]
2. Zhang, M.; Zhang, T.; Yan, C.; Zhang, Y. Driver Attention Prediction Based on Adaptive Fusion of Cross-Modal Features. *Appl. Sci.* **2026**, *16*, 2150. [CrossRef]
3. Fang, J.; Yan, D.; Qiao, J.; Xue, J.; Yu, H. DADA: Driver Attention Prediction in Driving Accident Scenarios. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 4959–4971. [CrossRef]
4. Xu, C.; Liu, H.; Li, Q.; Su, Y. Driver's Visual Fixation Attention Prediction in Dynamic Scenes Using Hybrid Neural Networks. *Digit. Signal Process.* **2023**, *142*, 104217. [CrossRef]
5. Itti, L.; Koch, C.; Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
6. Palazzi, A.; Abati, D.; Calderara, S.; Solera, F.; Cucchiara, R. DR (eye)VE: A Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 54–60. [CrossRef]
7. Cong, R.; Lei, J.; Fu, H.; Cheng, M.-M.; Lin, W.; Huang, Q. Review of Visual Saliency Detection with Comprehensive Information. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 2941–2959. [CrossRef]
8. Palazzi, A.; Abati, D.; Solera, F.; Cucchiara, R. Predicting the Driver's Focus of Attention: The DR (eye)VE Project. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1720–1733. [CrossRef] [PubMed]
9. Deng, T.; Yan, H.; Qin, L.; Ngo, T.; Manjunath, B.S. How Do Drivers Allocate Their Potential Attention? Driving Fixation Prediction via Convolutional Neural Networks. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 2146–2154. [CrossRef]
10. Chen, Y.; Nan, Z.; Xiang, T. FBLNet: FeedBack Loop Network for Driver Attention Prediction. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 13371–13380. [CrossRef]
11. Hu, Z.; Lv, C.; Hang, P.; Huang, C.; Xing, Y. Data-Driven Estimation of Driver Attention Using Calibration-Free Eye Gaze and Scene Features. *IEEE Trans. Ind. Electron.* **2022**, *69*, 1800–1808. [CrossRef]
12. Zhang, Y.; Chen, Y.; Gao, C. Deep Unsupervised Multi-Modal Fusion Network for Detecting Driver Distraction. *Neurocomputing* **2021**, *421*, 26–38. [CrossRef]

13. Gou, C.; Zhou, Y.; Li, D. Driver Attention Prediction Based on Convolution and Transformers. *J. Supercomput.* **2022**, *78*, 8268–8284. [CrossRef]
14. Huang, T.; Fu, R. Prediction of the Driver's Focus of Attention Based on Feature Visualization of a Deep Autonomous Driving Model. *Knowl. Based Syst.* **2022**, *247*, 109006. [CrossRef]
15. He, K.; Girshick, R.; Dollar, P. Rethinking ImageNet Pre-Training. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4918–4927. [CrossRef]
16. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef] [PubMed]
17. Kang, B.; Lee, Y. High-Resolution Neural Network for Driver Visual Attention Prediction. *Sensors* **2020**, *20*, 2030. [CrossRef] [PubMed]
18. Kang, B.; Lee, Y. A Driver's Visual Attention Prediction Using Optical Flow. *Sensors* **2021**, *21*, 3722. [CrossRef] [PubMed]
19. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239. [CrossRef]
21. Yang, L.; Wang, S.; Teng, S. Panoptic Image Segmentation Method Based on Dynamic Instance Query. *Sensors* **2025**, *25*, 2919. [CrossRef] [PubMed]
22. Li, C.; Zou, Q.; Li, G.; Zhang, J.; Zhang, Y. Landslide Segmentation in High-Resolution Remote Sensing Images: The Van-UPerAttnSeg Framework with Multi-Scale Feature Enhancement. *Remote Sens.* **2025**, *17*, 1265. [CrossRef]
23. Baee, S.; Pakdamanian, E.; Kim, I.; Feng, L.; Ordonez, V.; Barnes, L. MEDIRL: Predicting the Visual Attention of Drivers via Maximum Entropy Deep Inverse Reinforcement Learning. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 13178–13188. [CrossRef]
24. Xia, Y.; Zhang, D.; Kim, J.; Nakayama, K.; Zipser, K.; Whitney, D. Predicting Driver Attention in Critical Situations. In *Proceedings of Computer Vision Systems; Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2019; Volume 11754, pp. 658–674. [CrossRef]
25. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.
26. Xu, C.; Jiang, B.; Su, Y. TransConvNet: Perform Perceptually Relevant Driver's Visual Attention Predictions. *Comput. Electr. Eng.* **2024**, *115*, 109104. [CrossRef]
27. Fu, R.; Huang, T.; Li, M.; Sun, Q. A Multimodal Deep Neural Network for Prediction of the Driver's Focus of Attention Based on Anthropomorphic Attention Mechanism and Prior Knowledge. *Expert Syst. Appl.* **2023**, *214*, 119157. [CrossRef]
28. Kornblith, S.; Shlens, J.; Le, Q.V. Do Better ImageNet Models Transfer Better? In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2656–2666. [CrossRef]
29. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How Transferable Are Features in Deep Neural Networks? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 3320–3328.
30. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
31. Zhou, Y.; Gou, C.; Guo, Z.; Cheng, Y.; Chang, H.J. Behavior-Aware Knowledge-Embedded Model for Driver Attention Prediction. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 10199–10212. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.