Driver Intention Prediction Based on Multi-Dimensional Cross-Modality Information Interaction

Mengfan Xue
Hangzhou Dianzi University
Jiannan Zheng
Hangzhou Dianzi University
Li Tao
Hangzhou Dianzi University
Yuerong Wang
Hangzhou Dianzi University
Dongliang Peng
dlpeng@hdu.edu.cn
Hangzhou Dianzi University

Research Article

Keywords: driver intention prediction, self-driving, multimodal learning, contrastive learning, deep neural networks

Posted Date: May 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2942479/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on March 15th, 2024.
See the published version at https://doi.org/10.1007/s00530-024-01282-3.

Driver Intention Prediction Based on Multi-Dimensional Cross-Modality Information Interaction

Mengfan Xue[1,#], Jiannan Zheng[1,#], Tao Li[1], Yuerong Wang[1],Dongliang Peng[1,*]

[1]School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang, 310018, China

[*]Corresponding author: dlpeng@hdu.edu.cn

[#]These authors contributed equally to this work and should be considered co-first authors

Abstract Driver intention prediction allow drivers to perceive possible dangers in the fastest time and has become one of the most important research topics in the field of self-driving in recent years. In this study, we propose a driver intention prediction method based on multi-dimensional cross-modality information interaction. First, an efficient video recognition network is designed to extract channel-temporal features of in-side (driver) and out-side (road) videos respectively, in which we design a cross-modality channel-spatial weight mechanism to achieve information interaction between the two feature extraction networks corresponding respectively to the two modalities, and we also introduce a contrastive learning module by which we force the two feature extraction networks to enhance structural knowledge interaction. Then, the obtained representations of in- and out-side videos are fused using a Res-Layer based module to get a preliminary prediction which is then corrected by incorporating the GPS information to obtain a final decision. Besides, we employ a multi-task framework to train the entire network. We validate the proposed method on the public dataset Brain4Car, and the results show that the proposed method achieves competitive results in accuracy while balancing performance and computation.

Keywords driver intention prediction · self-driving · multimodal learning · contrastive learning · deep neural networks

1. Introduction

Based on a report by the World Health Organization (WHO), fatal injuries resulting from road traffic accidents account for approximately 1.35 million deaths worldwide annually, with non-fatal incidents excluded [1]. Driver misconduct, which includes dangerous driving behaviors such as illegal lane changes or turns, speeding, and fatigue driving, has been found to be the primary cause of most traffic accidents according to research in the field of road safety [2]. In recent years, advanced driver assistance systems (ADAS) have been widely valued as a means to improve driving safety and prevent car accidents. Driver intention prediction, a key component of such systems, enables drivers to rapidly detect potential hazards and develop a range of solutions to enhance road safety [1,3,4].

In real-world situations, humans have a reaction time of approximately 2-3 seconds to respond to accidents. Therefore, driver intention prediction algorithms must be designed with sufficient expected

anticipation to aid drivers in making timely decisions [5,6]. Jain et al. released a dataset of natural driving containing in-vehicle and out-vehicle videos, GPS, and speed information, and achieved an accuracy of 86% with an expected anticipation of 3.5 seconds using a deep learning sensory fusion architecture [7]. In recent years, numerous studies have successfully achieved effective driver intention prediction based on the Brain4cars dataset. However, some issues still require further examination. Some research teams have primarily relied on in-side video and manually coded road information to predict driver maneuvers [8-10]. These studies have shown that driver behavior can ensure safe take-over behavior in conditionally autonomous driving [11]. Nevertheless, these works lack comprehensive information without further processing the out-side videos. The effectiveness of machine learning in intent recognition has been established [12-14]. Kim et al. augmented the in- and outside information using an artificial neural network (ANN) model and fed the augmented information to a support vector machine (SVM) to detect the driver's intention [15]. However, modeling temporal information using ANN is challenging, leading to suboptimal performance in intent prediction.

Deep learning has recently gained extensive development and application in various fields. Due to the principle of using a large number of neurons to simulate human perception, thinking, and other activities, researchers have employed deep learning to address driver intention prediction, with promising outcomes [1,10,16-18]. Generally, deep learning-based driver intention prediction approaches offer advantages such as automatic feature learning, end-to-end learning, and more comprehensive and superior performance of the learned features. However, most studies utilize 3D Conv [10,16,19], optical flow [1], or stacking of LSTM [3,16-18,20] to model the temporal information of video sequences, leading to issues of large network parameters and high algorithm deployment costs. Moreover, it is noteworthy that most studies neglect or do not effectively utilize GPS information.

This paper presents a novel driver intention prediction method that leverages multi-dimensional cross-modality information interaction. The proposed method uses an efficient video sequence feature extraction network to extract channel-temporal features from the in-side (driver) and out-side (road) videos. To balance deployment cost with high prediction accuracy, the method uses both long-term and short-term temporal modules to model global and local temporal information, respectively, and a channel attention module to extract channel information. The feature extraction process is enhanced with a cross-modality channel-spatial weight mechanism to achieve information interaction between the two modalities. Additionally, a contrastive learning module is introduced to force the two feature extraction networks to enhance structural knowledge interaction. The obtained representations from the in- and out-side videos are fused using a Res-Layer based module to obtain a preliminary prediction, which is further corrected by incorporating GPS information to obtain a final decision. To ensure high-precision intention prediction while performing efficient single-modality feature extraction, a multi-task framework is employed to train the entire network. The contributions of this paper include the proposed cross-modality information interaction, the efficient video sequence feature extraction network, and the multi-task framework for high-precision intention prediction. The key contributions in

this paper can be highlighted as follows:

(1) We propose a driver intention prediction method based on multi-dimensional cross-modality information interaction.

(2) We design an efficient video sequence feature extraction network to extract channel-temporal features of in-side (driver) and out-side (road) videos respectively, in which we design a cross-modality channel-spatial weight mechanism to achieve information interaction between the two feature extraction networks corresponding respectively to the two modalities, and we also introduce a contrastive learning module by which we force the two feature extraction networks to enhance structural knowledge interaction.

(3) We propose a prediction module based on Res-Layer and GRU classifier to get a preliminary prediction which is then corrected by incorporating the GPS information to obtain a final decision.

2. Related Work

This section mainly sorts out the related work of the proposed method and reviews the technical and algorithm of driver intention prediction from the following three perspectives, including (1) the application of Convolutional Neural Network (CNN) [21] in driver intention prediction; (2) video sequence analysis; and (3) cross-modal information interaction.

2.1 Driver Intention Prediction by CNN

Convolutional neural networks (CNN) have shown promising results in driver intention prediction tasks due to their ability to extract and learn high-level representations of image and frame sequence features, mimicking the human visual system. For example, Bonyani et al. employed a DenseNet121 architecture with Dropout and Avg-Pooling techniques, as well as LSTM and Global attention modules, and used RAFT and FlowNet2 to extract optical flow of in-side and out-side video sequences [1]. Some researchers have utilized 3D Conv to model the temporal information of video frame sequences. Chen et al. proposed a two-stream structure based on a deep three-dimensional CNN [16], while Rong et al. used 3D-ResNet to extract video spatiotemporal features [10]. Recent studies have demonstrated that CNN-based algorithms achieve state-of-the-art (SOTA) performance on datasets such as Brain4Car [1].

2.2 Video Sequence Analysis

Temporal modeling plays a crucial role in video sequence analysis. While 3D Conv-based networks have been widely used for this purpose, they are computationally expensive. To achieve efficient global temporal information representation, TSM [26] first uses a temporal shift algorithm based on a 2D Conv network for temporal modeling, which addresses the issue that traditional 2D Conv cannot capture the relationship in the temporal dimension, and the high deployment cost of 3D Conv-based methods. Similarly, to reduce the computational pressure caused by optical flow extraction [27-29], some research teams propose to use 2D Conv-based modules to extract difference information instead of optical flow. To balance the deployment cost of the algorithm while predicting driver intention with high accuracy, we propose a long-term and short-term temporal module to model the

global and local temporal information of the video sequence respectively. Specifically, our long-term temporal module adopts a different time-shifting method from TSM to achieve long-term temporal information modeling with low computational cost. To efficiently represent short-term temporal information, our short-term temporal module refers to the motion excitation (ME) module [30]. Unlike the ME module, our short-term temporal module calculates the feature-level frame differences of the current frame and two adjacent frames to represent the motion information more comprehensively.

An SE block [31], which models the dependencies of each channel to selectively strengthen useful features and suppress useless ones through global information, has been proven to improve the channel information representation ability of a network in image classification tasks. Recent studies have shown that inserting SE blocks into a network can significantly improve its performance [32–34]. For video recognition tasks, Wang et al. proposed the addition of channel excitation to a video recognition network to enhance the channel information representation ability and improve the overall performance [35,36]. Building upon these findings, we propose a channel attention module that is structurally similar to an SE block [31]. We replace the fully connected layer in the SE block with a 2D Conv layer and add a 1D Conv layer to improve the channel information modeling ability.

Based on the above work, we design an efficient video sequence feature extraction network, which utilizes the channel attention module and long- and short-term temporal module, to extract channel-temporal features of in-side (driver) and out-side (road) videos respectively.

2.3 Cross-Modal Information Interaction

The task of driver intension prediction is a challenging one that requires the integration of multiple types of information to achieve accurate results. Prior methods [8–10] have focused primarily on using in-side video and manually coded road information to predict driver maneuvers, which limits the ability to obtain comprehensive multimodal information. To address this, many research teams have emphasized the importance of cross-modal information interaction and designed effective multi-modal fusion methods. However, these methods typically focus on multimodal fusion in a single dimension (i.e., either feature extractor or classifier) [1,10,16–18], and often overlook the potential benefits of incorporating GPS information. This is a noteworthy limitation in the field that warrants further exploration.

In order to achieve effective multimodal information fusion, we propose a driver intention prediction method based on multi-dimensional cross-modality information interaction. In the feature extraction process of in-side and out-side videos, we first design a cross-modality channel-spatial weight mechanism to achieve information interaction between the two feature extraction networks corresponding respectively to the two modalities. And to measure the structural knowledge between cross-modal and cross-sample feature vector representations, we introduce a contrastive learning module by which we force the two feature extraction networks to enhance structural knowledge interaction. Then, the obtained representations of in- and out-side videos are fused using a Res-Layer based module to get a preliminary prediction, which is then corrected by incorporating the GPS

information to obtain a final decision. More specifically, we design a 1D deep neural network based on the residual structure [37] to extract multimodal fusion features, and use a GRU classifier to predict a preliminary result, and finally, we use the lane information and intersection information in the GPS information to correct the preliminary prediction results to get the final decision.

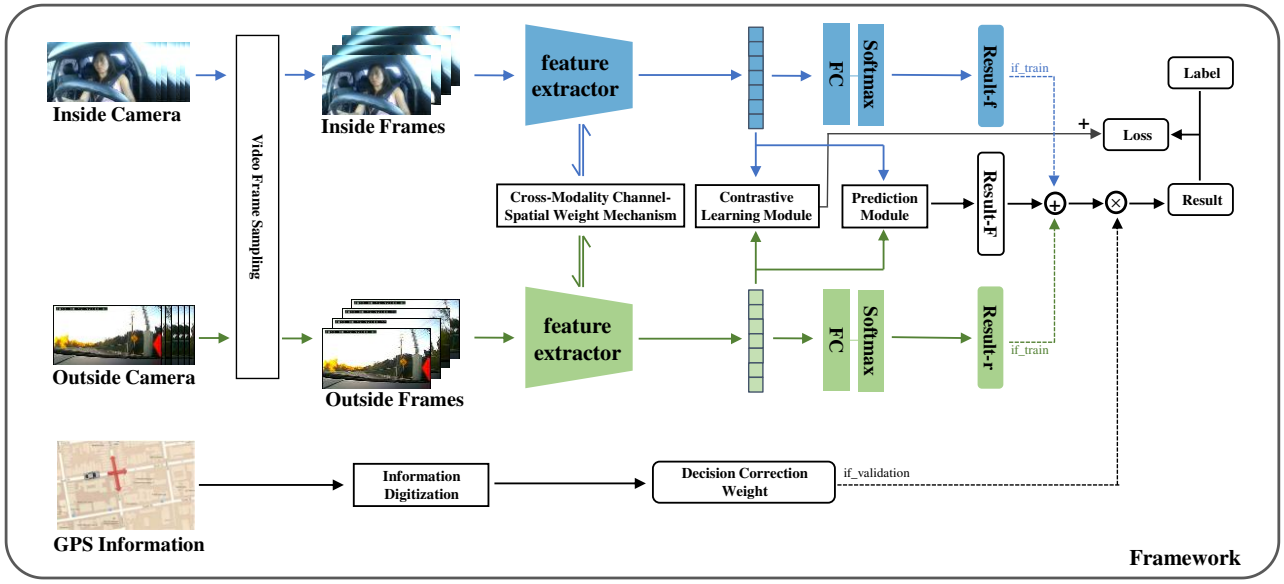

Figure 1. Framework of the proposed driver intention prediction method.

3. Our Method Design

Our proposed driver intention prediction framework diagram is shown in Figure 1. In this section, we introduce the technical details of the proposed driver intention prediction method. First, an efficient video recognition network is designed to extract channel-temporal features of in- and out-side videos respectively, in which we design a cross-modality channel-spatial weight mechanism to achieve information interaction between the two feature extraction networks corresponding respectively to the two modalities, and we also introduce a contrastive learning module by which we force the two feature extraction networks to enhance structural knowledge interaction. Then, the obtained representations of in- and out-side videos are fused using a Res-Layer based module to get a preliminary prediction which is then corrected by incorporating the GPS information to obtain a final decision. Besides, we employ a multi-task framework to train the entire network. We adopt the frame sampling strategy proposed by TSN [38], divide the input video V (in-side and out-side videos) into T video segments at equal intervals, namely, $V=(V_1,V_2,\cdots V_T)$, and randomly extract a frame from each video segment to obtain T frames of images as the input sequence.

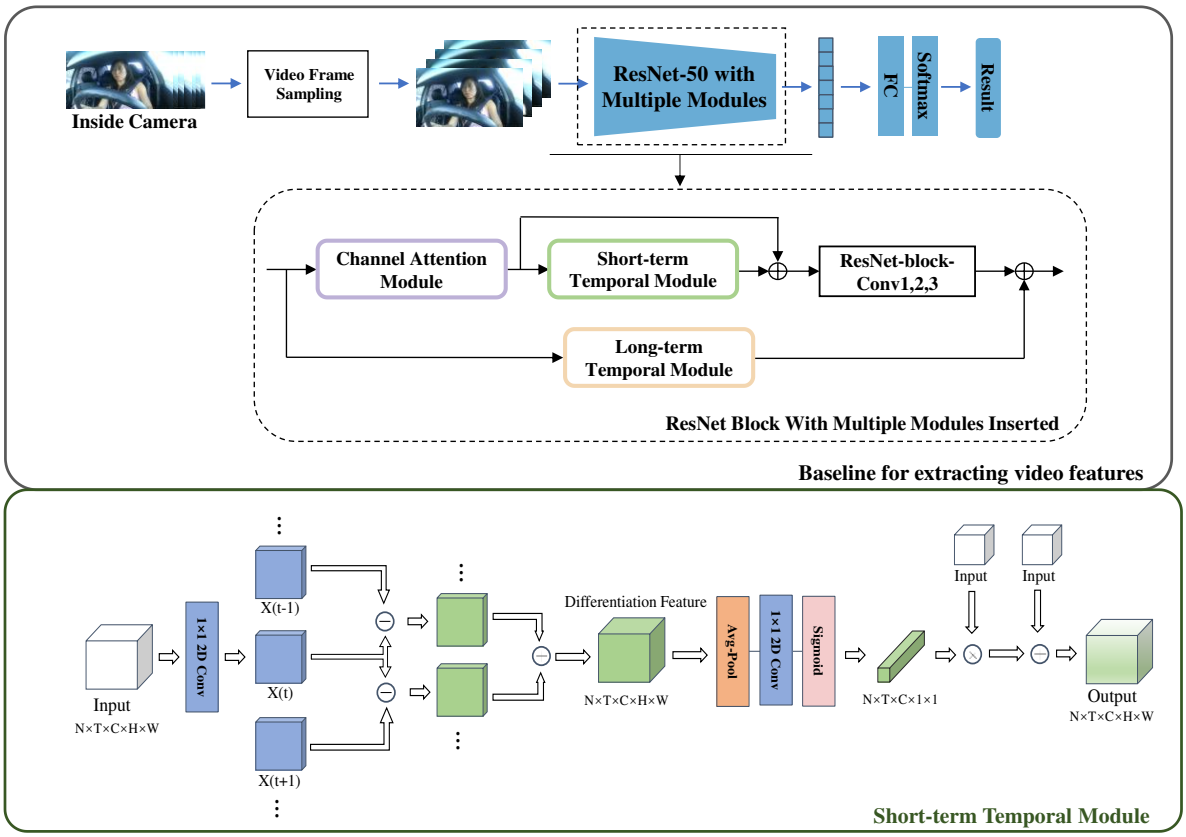

Figure 2. Baseline for extracting video features (top) and implementation of the short-term temporal module (bottom).

3.1 Baseline for extracting video features

The modeling of channel and temporal information is crucial for effective video sequence analysis [36]. The proposed method presents an efficient video sequence feature extraction network for extracting channel-temporal features from in-side (driver) and out-side (road) videos separately. As depicted in Figure 2 (top), we employ a long-term and short-term temporal module to capture global and local temporal information of the video sequence, respectively, and a channel attention module to extract channel information. We have provided a more detailed explanation and further discussion on the feature weighting operation of the proposed module in *Section 5.2*. Then, the designed multimodule is incorporated into each ResNet block of the 2D ResNet-50 [37] in a logical manner.

More specifically, the structure of the channel attention module is similar to that of an SE block [31]. We refer to the ACTION module [34] to add a 1D Conv between two FC layers of an SE block to strengthen the channel information modeling ability, and the FC layers in the SE block are replaced by a 2D Conv layer. Given an input of $X \in \mathbb{R}^{N\times T\times C\times H\times W}$, where N is the batch size, C is the number of channels, and H、W are the height and width of the input image, we utilize the dependencies of each channel obtained by modeling the above steps to strengthen the features containing useful information and suppress useless features, which can be interpreted as:

$$X_O = X_I + X_I \square W_C, \quad X_I \in \square^{N\times T\times C\times H\times W}, W_C \in \square^{N\times T\times C\times 1\times 1} \quad (1)$$

Where X_I and X_O represent the input and output of the channel attention module, and W_C represents the channel weight calculated by the proposed method.

The long-term temporal module adopts a similar idea to the temporal shift module(TSM) [26]. For a given input $X_I \in \square^{N\times T\times C\times H\times W}$, X_I is divided into 8 equal parts according to the channel dimension, namely, $X_I=[X_1,X_2,\cdots,X_8]$, where $X_n \in \square^{N\times T\times \frac{C}{8}\times H\times W}$, $n=1,2,\cdots,8$. X_1 and X_3 are shifted forward in the temporal dimension, and X_2 is shifted backward in the temporal dimension. After the feature is moved, "outlier features" appear in the direction of temporal movement, and "vacant features" appear in the opposite direction of the temporal movement. We transfer the "outlier features" to the "vacant features" and restore the feature to the size before the temporal change.

To efficiently represent local temporal information, the short-term temporal module refers to the ME module [30]. As shown in Figure 2 (bottom), unlike the ME module, to extract the motion information more fully, the short-term temporal module first calculates the feature-level frame differences of the current frame and two adjacent frames to represent the motion information so the network can automatically capture the difference information between adjacent frames. Then, similar to the channel attention module structure, we utilize the obtained motion information to enhance motion-sensitive features, which can be interpreted as:

$$X_O = X_I + X_I \square W_S, \quad X_O \in \square^{N\times T\times C\times H\times W}, X_I \in \square^{N\times T\times C\times 1\times 1}, W_S \in \square^{N\times T\times C\times 1\times 1} \quad (2)$$

Where X_I and X_O represent the input and output of the short-term temporal module, and W_S represents the motion information weight calculated by the proposed module.

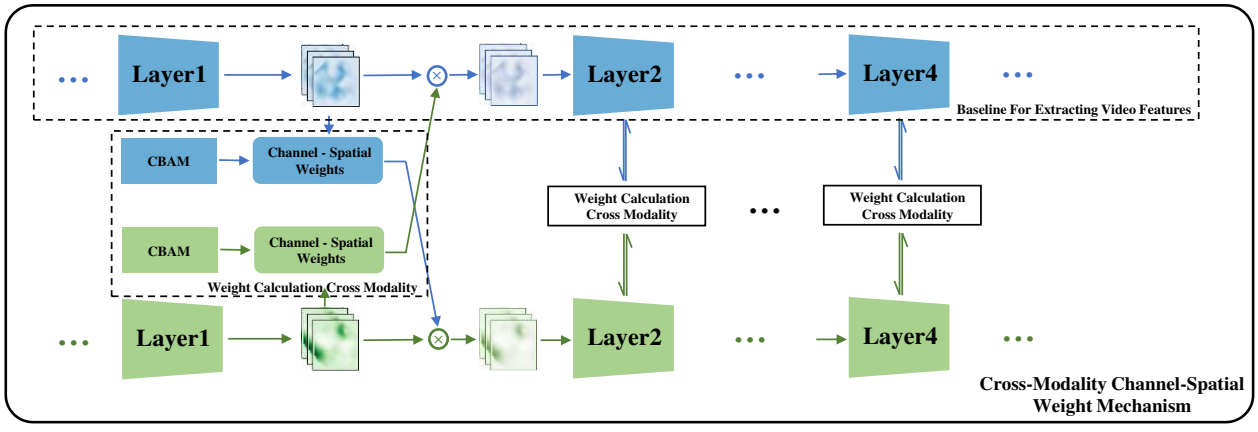

Figure 3. Implementation of the cross-modality channel-spatial weight mechanism.

3.2 Cross-Modality Channel-Spatial Weight Mechanism

The effective interaction of multimodal information during feature extraction is critical for successful cross-modal fusion. However, previous studies have often overlooked this aspect [1,10,16–

18]. In this paper, we propose a novel information interaction mechanism that avoids the performance limitations of simply concatenating features or increasing the computational load of the network through excessive use of two-dimensional convolutional layer stacks. Specifically, we employ an attention mechanism [31,39] that models dependencies between input data and enhances useful features. We draw inspiration from the Convolutional Block Attention Module (CBAM) proposed by Woo et al. [40], which sequentially infers attention maps along two independent dimensions (i.e., channel and spatial), and multiplies the resulting attention maps with the input feature map for adaptive feature refinement.

In order to achieve effective information complementation in the process of in-side and out-side video feature extraction, as shown in Figure 3, we propose a cross-modality channel-spatial weight mechanism. We first use CBAM [40] to extract the channel and spatial dependencies of the features of in- and out-side video sequences, then apply the attention weight (channel and spatial) of the in-side video features to the out-side video features, and vice versa, the attention weight of the out-side video feature is applied to the in-side video features, which can be interpreted as:

$$\begin{cases} X^{face}{}_{O} = X^{face}{}_{I} \otimes W^{road}{}_{C} + X^{face}{}_{I} \otimes W^{road}{}_{S} \\ X^{road}{}_{O} = X^{road}{}_{I} \otimes W^{face}{}_{C} + X^{road}{}_{I} \otimes W^{face}{}_{S} \end{cases} \quad (3)$$

Where $X^{face}{}_{I} \in \mathbb{R}^{N\times T\times C\times H\times W}$ and $X^{road}{}_{I} \in \mathbb{R}^{N\times T\times C\times H\times W}$ represent the in-side(driver) and out-side(road) video sequence features to be weighted, $X^{face}{}_{I} \in \mathbb{R}^{N\times T\times C\times H\times W}$ and $X^{road}{}_{I} \in \mathbb{R}^{N\times T\times C\times H\times W}$ represent the weighted in- and out-side video sequence features, $W^{road}{}_{C} \in \mathbb{R}^{N\times T\times C\times 1\times 1}$ and $W^{road}{}_{S} \in \mathbb{R}^{N\times T\times 1\times H\times W}$ represent channel attention weights and spatial attention weights generated by in-side video features, $W^{face}{}_{C} \in \mathbb{R}^{N\times T\times C\times 1\times 1}$ and $W^{face}{}_{S} \in \mathbb{R}^{N\times T\times 1\times H\times W}$ represent channel attention weights and spatial attention weights generated by in-side video features.

We employ the proposed cross-modality channel-spatial weight mechanism to efficiently complement multi-modal information. It is important to highlight that this cross-modal feature interaction is not limited to a single layer, but instead, weight calculation and cross-modal feature weighting are performed at the output features of each block of the two ResNet-50 networks (with channel and temporal information modeling) corresponding to the two modalities.

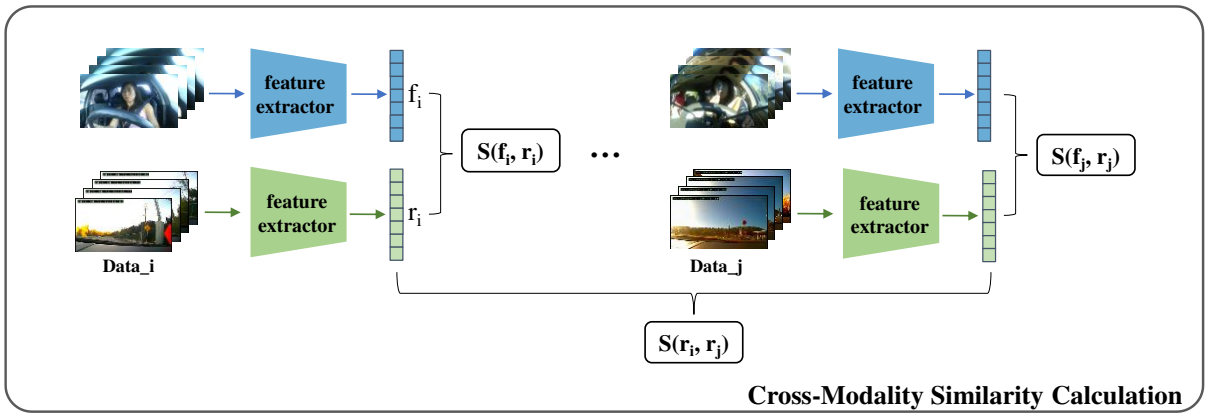

Figure 4. Implementation of the contrastive learning module.

3.3 Contrastive Learning Module

Contrastive learning is a kind of representation learning, and its main idea is to make the representations of similar samples close, while dissimilar ones are far away [41–43]. In recent years, the idea of contrastive learning has begun to be exploited in the field of multimodal learning [44,45], specifically, multimodal data of the same sample are regarded as positive samples with similar structures, while data of different samples are regarded as negative samples that need to be kept away. Inspired by these studies, as shown as Figure 4, the proposed algorithm designs a contrastive learning module to force the network to learn structural knowledge interaction of multimodal data. In order to bring the positive sample features closer and the negative sample features farther away, we use the Cosine Similarity to measure the distance represented by the vector. The Cosine Similarity between x and y, $S(x, y)$, can be expressed as:

$$S(x, y) = \frac{x \cdot y}{\| x \| \times \| y \|} = \frac{\sum_{i=1}^{n} (x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i^{\,2}} \times \sqrt{\sum_{i=1}^{n} y_i^{\,2}}} \quad (4)$$

Then, for a pair of videos (in- and out-side) belonging to the same sample, we believe that their feature representations should have structural similarity, while feature representations belonging to different samples should have a certain distance. By calculating the Cosine Similarity between the features of positive and negative samples, the similarity penalty p_s can be obtained, which can be expressed as:

$$p_s = -\log\left(\frac{e^{S(f_i, r_i)/\tau}}{\sum_{i \neq j} \left(e^{S(f_i, r_j)/\tau} + e^{S(f_i, f_j)/\tau} + e^{S(r_i, r_j)/\tau} + e^{S(r_i, f_j)/\tau} \right)} \right) \quad (5)$$

Where f_i and r_i represent the feature vector representations of the in-side (driver) and out-side (road) videos of the i-th sample respectively, τ is a temperature parameter that controls the concentration level of the distribution [43].

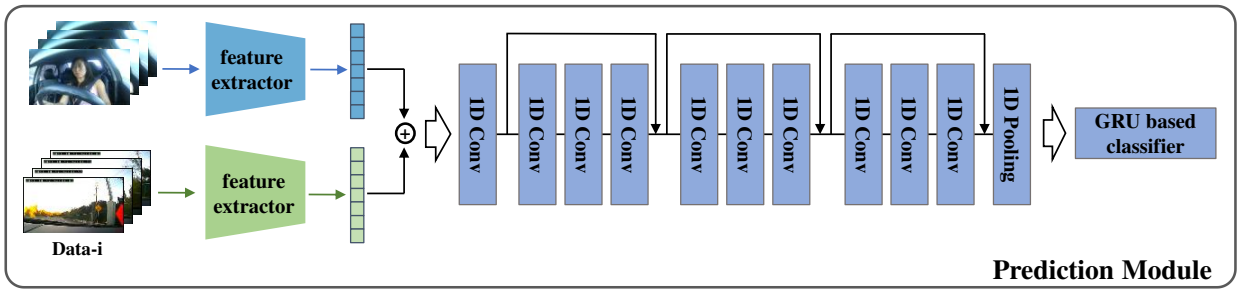

Figure 5. Implementation of the prediction module.

3.4 Prediction Module and Decision Correction

The design of an effective classifier is crucial for the accuracy of the prediction algorithm. However, for video recognition tasks, simply taking the average of frame-wise predictions as the output of the classifier is not sufficient to model the temporal information of the feature sequence [26,30,36]. In contrast, many recent studies have shown that building a fusion network based on 1D Conv is an effective approach for feature fusion of multimodal data [1,10,16]. Furthermore, it is noteworthy that previous studies often overlook or underutilize the valuable GPS information.

In order to build an efficient classifier, as shown as Figure 5, we propose a prediction module based on Res-Layer [37] and GRU [46] classifier to get a preliminary prediction which is then corrected by incorporating the GPS information to obtain a final decision. We employ a simple but effective approach to utilize GPS information for decision correction. For example, if the driver to be predicted is in the leftmost lane, we think it is extremely unlikely that the driver will change lanes to the left, so we suppress the prediction score of "left change lane ". For the categories [go straight, left change lane, turn left, right change lane, turn right], the process of decision correction can be expressed as:

$$\omega_d = \begin{cases} \text{if-train:}[1,1,1,1,1]^T \\ \text{if-validation:} \begin{cases} [1,\lambda,1,1,1]^T, \text{no left change lane} \\ [1,1,1,\lambda,1]^T, \text{no right change lane} \\ [1,1,\lambda,1,\lambda]^T, \text{no turn} \\ [1,1,1,1,1]^T, \text{other} \end{cases} \end{cases} \tag{6}$$

Where λ represent the suppression coefficient to suppress the score of the "wrong choice", and ω_d represents the correction weight acting on the preliminary prediction.

3.5 Multi-Task Framework

To ensure accurate driver intention prediction without compromising the efficiency of single-modality feature extraction, we utilize a multi-task framework to train the network. The primary task is to predict driver intention using the prediction module, while the output of the single-modal network serves as an auxiliary task to improve feature extraction for each modality and enhance the accuracy of

prediction. Based on the above multi-task and multi-modal framework, the loss function of the proposed method can be expressed as:

$$L=E\{\omega_d[(1-\alpha)x_F+\frac{\alpha(x_f+x_r)}{2}],y\}+p_s \quad (7)$$

Where $E(x,y)$ Cross entropy loss, x_F represent the output of the prediction module, x_f and x_r represent the output of the in-side (driver) and out-side (road) single-modal network respectively, and α represent the balance factor for multi-task learning.

4. Experiments

This section presents the experimental details and results of the proposed driver intention prediction method that utilizes multi-dimensional cross-modality information interaction. The performance of the proposed method on the driver intention prediction task is evaluated using the public Brain4cars dataset, and its recognition accuracy is compared against other algorithms. To assess the effectiveness of the proposed approach, several targeted ablation experiments are conducted to evaluate the contribution of various components, including the baseline for video feature extraction, the cross-modality channel spatial weight mechanism, the comparative learning module, the prediction module, and decision correction. The evaluation metrics used in the experiments include accuracy and F1-score, which are standard metrics for classification tasks. The details of the dataset and experimental setup are provided in *Section 4.1* and *Section 4.2*, while the results and analysis are presented in *Section 4.3* and *Section 4.4*.

4.1 Dataset

The Brain4cars dataset is a publicly available natural driving dataset, which consists of in-vehicle and out-vehicle videos, GPS, and speed information. It was released by Jain et al., who achieved an accuracy of 86% in expected anticipation of 3.5 seconds using a deep learning sensory fusion architecture [7]. Since human response time to an accident is typically 2-3 seconds in practical situations, it is crucial for driver intention prediction algorithms to have sufficient anticipation to enable drivers to make real-time decisions [5,6]. Therefore, we evaluated the performance of our proposed driver intention prediction method on the Brain4cars dataset to verify its effectiveness. In addition, we compared the recognition accuracy of our proposed method with other state-of-the-art algorithms, and conducted targeted ablation experiments to validate the effectiveness of different components of our method, including the baseline for extracting video features, the cross-modality channel spatial weight mechanism, the comparative learning module, the prediction module, and decision correction.

4.2 Implementation Details

The proposed driver intention prediction method is a multimodal and multi-task network that utilizes in-side and out-side video, GPS, and speed information as multimodal inputs. The video sequences are the primary modality, and ResNet-50 pre-trained on the ImageNet dataset is used as the backbone network for feature extraction. Multiple modules are inserted into each ResNet block of ResNet-50 in a reasonable manner to enhance the feature extraction process. For pre-training, we use the EgoGesture [47] dataset, a large-scale gesture recognition dataset. In the frame sampling strategy,

we adopt the same method as TSN[38], divide the input video V into T video segments at equal intervals, that is, $V=(V_1,V_2,\cdots V_T)$, and randomly extract one frame from each video segment to obtain T image frames as the input sequence (in our experiments, T=8). During training, we use corner cropping and scale-jittering as data augmentation. After cropping, the size of the frame sequence input to the recognition network is $N\times T\times C\times 224\times 224$, where N is the batch size, T is the number of segments, and C is the frame image channel number.

GPS information is used as an auxiliary modality to correct the preliminary prediction results based on the analysis of video sequences. We digitize GPS information and record it as three digital features, including the current lane, total number of lanes, and intersection information. The current lane is counted from right to left, and the intersection information indicates the presence or absence of intersections near the vehicle using '1' and '0,' respectively. K-fold Cross Validation with K=5 is utilized to address the limited amount of data available in our experiments.

Table 1. Comparisons with the state-of-the-art. We compare the proposed driver intention prediction method based on multi-dimensional cross-modality information interaction with other state-of-the-art methods on the Brain4cars datasets[7].

Method	Data Source	Fold1		Fold2		Fold3		Fold4		Fold5		Mean+std	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
ConvLSTM auto-encoder & ResNet50	In-out	-	-	-	-	-	-	-	-	-	-	84.0±0.1	84.3±0.1
RNN-LSTM	In-out	-	-	-	-	-	-	-	-	-	-	92.1	86.1
STEDII-GRU	In-out	-	-	-	-	-	-	-	-	-	-	92.1±1.9	90.0±2.2
CF-LSTM and predictive-Bi-LSTM-CRF	In-out	-	-	-	-	-	-	-	-	-	-	92.4	93.6
CNN-LSTM	In-out	-	-	-	-	-	-	-	-	-	-	94.1	-
F-RNN-DMT	Inside	-	-	-	-	-	-	-	-	-	-	96.2	94.11
DIPNet	In-out	97.8	-	97.8	-	98.9	-	98.9	-	98.9	-	**98.5±0.6**	**98.9**
Ours	In-out	94.8	94.0	96.6	94.9	94.9	93.6	95.0	94.3	96.6	96.2	95.6±0.6	94.5±0.9

4.3 Comparisons with the state-of-the-art

Table 1 presents a comparison of the proposed driver intention prediction method, based on multi-dimensional cross-modality information interaction, with other state-of-the-art methods on the Brain4cars dataset [7]. Recently, several studies have used deep learning to address the problem of driver intention prediction and have achieved promising results. However, most of these studies rely on 3D Conv (STEDII-GRU[16], DIPNet[19]) or stacking of LSTM (ConvLSTM auto-encoder & ResNet50[10], RNN-LSTM[17], CF-LSTM and predictive-Bi-LSTM-CRF[20], CNN-LSTM[15], F-RNN-DMT[3]) to model the temporal information of video sequences. This approach has resulted in large network parameters and high algorithm deployment costs.

In contrast, the proposed driver intention prediction method achieves highly competitive results

while balancing the algorithm deployment costs (Acc=95.6±0.6, F1=94.5±0.9), which is slightly lower than the current state-of-the-art method DIPNet [19] (Acc ↓ 3.9%, F1 ↓ 4.4%). However, the proposed method efficiently extracts channel-temporal features of in-side (driver) and out-side (road) videos, thereby achieving high-precision driver intent prediction while balancing the deployment cost of the algorithm. We further discuss and compare the computational cost and runtime of the algorithm in *Section 5.1*.

4.4 Ablation Study

In this section, we demonstrate the ablation experiments we design from four perspectives to show the superiority of the proposed driver intention prediction method in the structural design, including 1) the superiority of the baseline design for extracting video features, 2) the effectiveness of the cross-modality channel spatial weight mechanism, 3) the effectiveness of the comparative learning module, and 4) the effectiveness of the prediction module and the decision correction.

Table 2. The superiority of the baseline design for extracting video features.

Method	Data Source	Acc	F1
3D ResNet-50	In-out	80.3	84.5
2D ResNet-50 with TSN	In-out	82.2	79.3
The Proposed Video Classification Network	In-out	**95.6**	**94.5**

4.4.1 The superiority of the baseline design for extracting video features

Efficient video feature extraction is essential for accurate driver intention prediction, and the proposed method achieves this by designing an efficient video recognition network that extracts channel-temporal features of in-side (driver) and out-side (road) videos. In Table 2, the proposed video recognition network is compared with 3D ResNet-50 and 2D ResNet-50 with TSN on the Brain4cars dataset. The experimental results demonstrate that the proposed video recognition network outperforms 3D ResNet-50 with higher computational complexity (Acc ↑ 15.3%, F1 ↑ 10.0%), indicating that the proposed network can effectively model video sequence features even on a 2D Conv architecture. Moreover, the performance of the proposed video recognition network is also better than 2D ResNet-50 with TSN (Acc ↑ 13.4%, F1 ↑ 15.2%), which indicates that the proposed multiple modules, including the channel attention module and the long- and short-term temporal module, can efficiently extract channel-temporal information from videos while achieving high-performance video classification. These results confirm the effectiveness of the proposed baseline design for extracting video features.

Table 3. The effectiveness of the cross-modality channel spatial weight mechanism.

Layer1	Layer2	Layer3	Layer4	Acc	F1
				90.8	89.9

√				94.2	93.8
√	√			94.6	94.0
√	√	√		95.2	94.2
√	√	√	√	**95.6**	**94.5**

4.4.2 The effectiveness of the cross-modality channel spatial weight mechanism

The proposed cross-modality channel-spatial weight mechanism is a key feature of the proposed driver intention prediction method, which enables effective cross-modal information fusion during multimodal feature extraction. The mechanism calculates weights for each channel and spatial location in the output features of each block of the two ResNet-50 corresponding to the in-side and out-side videos, respectively. The weights are then used to perform cross-modal feature weighting, enabling effective information complementation between the two modalities. Experimental results show that the proposed mechanism significantly improves the algorithm performance, with an increase in accuracy and F1 score by 3.4~4.8% and 3.9~4.6%, respectively. Moreover, the results show that the more layers the mechanism is applied to, the better the recognition performance of the network. These findings suggest that the proposed mechanism is effective in facilitating cross-modal information fusion and improving the accuracy of driver intention prediction.

Table 4. The effectiveness of the comparative learning module.

	None	$p_s = -\log\left(e^{S(f_i, r_i)/\tau}\right)$	$p_s = -\log\left(\frac{e^{S(f_i, r_i)/\tau}}{\sum_{i \neq j}(\ldots)}\right)$
Acc	94.5	95.2	95.6
F1	93.8	94.5	94.5

4.4.3 The effectiveness of the comparative learning module

To achieve effective multimodal information fusion, it is crucial to have direct structural knowledge interaction of multimodal features. To address this, we propose a contrastive learning module that forces the network to learn the structural knowledge interaction of multimodal data. The module treats multimodal data of the same sample as positive samples with similar structures, while data of different samples are regarded as negative samples that need to be kept apart. The distance represented by the vector is measured using Cosine Similarity. In Table 4, we compare algorithm performance before and after applying the proposed contrastive learning module and using different similarity penalty calculation formulas. The experimental results demonstrate that applying the proposed contrastive learning module improves algorithm performance (Acc ↑ 0.7~1.1%, F1 ↑ 0.7%), indicating the module's efficacy for effective multimodal information interaction. To design a similarity

penalty that can enable effective structural knowledge interaction, we refer to the mainstream definition of comparative learning loss function [41–45]. We designed the proposed similarity penalty in *Section 3.3* and observed performance improvements (Acc ↑ 0.4%).

Table 5. The effectiveness of the prediction module.

	MLP	DenseLayer	ResLayer	ResLayer with GRU
Acc	90.3	88.8	95.1	95.6
F1	90.3	84.3	93.2	94.5

Table 6. The effectiveness of the decision correction.

	None	$\lambda = 0.5$	$\lambda = 0.1$
Acc	90.3	93.8	95.1
F1	90.3	93.3	93.2

4.4.4 The effectiveness of the prediction module and the decision correction

The classifier design is a crucial step in the driver intention prediction algorithm. To construct an efficient classifier, we propose a prediction module based on Res-Layer [37] and GRU [46] classifier. This module obtains a preliminary prediction, which is further refined by incorporating GPS information to obtain a final decision. In Table 5, we compare the proposed ResLayer with Denselayer and MLP structures, and the results show that the ResLayer structure outperforms the other two structures (Acc ↑ 4.8~6.3%, F1 ↑ 2.9~8.9%). Moreover, after incorporating GRU in the ResLayer, the algorithm performance was further improved (Acc ↑ 0.5%, F1 ↑ 1.3%), indicating that GRU can effectively model temporal information during the classification process.

In Table 6, we evaluate the proposed decision correction method with different suppression coefficients. The results show that decision correction significantly improves the classification accuracy (Acc ↑ 3.5~4.8%, F1 ↑ 2.9~3.0%). Furthermore, a smaller suppression coefficient ($\lambda = 0.1$) leads to better performance of the decision correction method (Acc ↑ 2.3%, F1 ↓ 0.1%), indirectly confirming the effectiveness of our proposed method.

5. Discussion

In this section, we will make necessary supplements and further discussions on some content beyond the main experiment , including 1) efficient driver intention prediction, 2) training visualization.

Table 7. Model complexity of the proposed method. We compare the model complexity with no module, a single module inserted, and the proposed video classification network.

	No module	channel attention module	long-term temporal module	short-term temporal module	Multiple modules
FLOPs	22.37 G	24.26 G	22.37 G	24.27 G	**26.17 G**
Params.	11.14 M	11.15 M	11.14 M	11.68 M	**11.69 M**

5.1 Efficient Driver Intention Prediction

Driver intention prediction is a crucial aspect of intelligent driving that has gained attention from the scientific community. However, previous studies have mainly focused on improving performance without considering the deployment cost of the algorithm on onboard devices [1,3,10,15–17,19-20]. In this study, we propose an efficient video sequence feature extraction network to extract channel-temporal features from in-side (driver) and out-side (road) videos separately. To model global and local temporal information of the video sequence, we use a long-term and short-term temporal module, respectively. Additionally, we employ a channel attention module to extract channel information. In *Section 4.4.1*, we have verified the effectiveness of the baseline design for extracting video features. In Table 7, we compare the model complexity of no module, a single module inserted, and the proposed video classification network. It can be observed that, there is only a small increase in model complexity after inserting a single module into the baseline (FLOPs ↑ 0~4.03% and Params. ↑ 0~4.85%). And our proposed network has a certain increase in model complexity (FLOPs ↑ 16.98% and Param. ↑ 4.94%), but considering the 13.4% improvement in accuracy on the Brain4cars dataset [7] compared to the backbone of the network (ResNet-50 with TSN), we believe that a small increase in computation cost is worthwhile to achieve a higher accuracy with reasonable deployment cost.

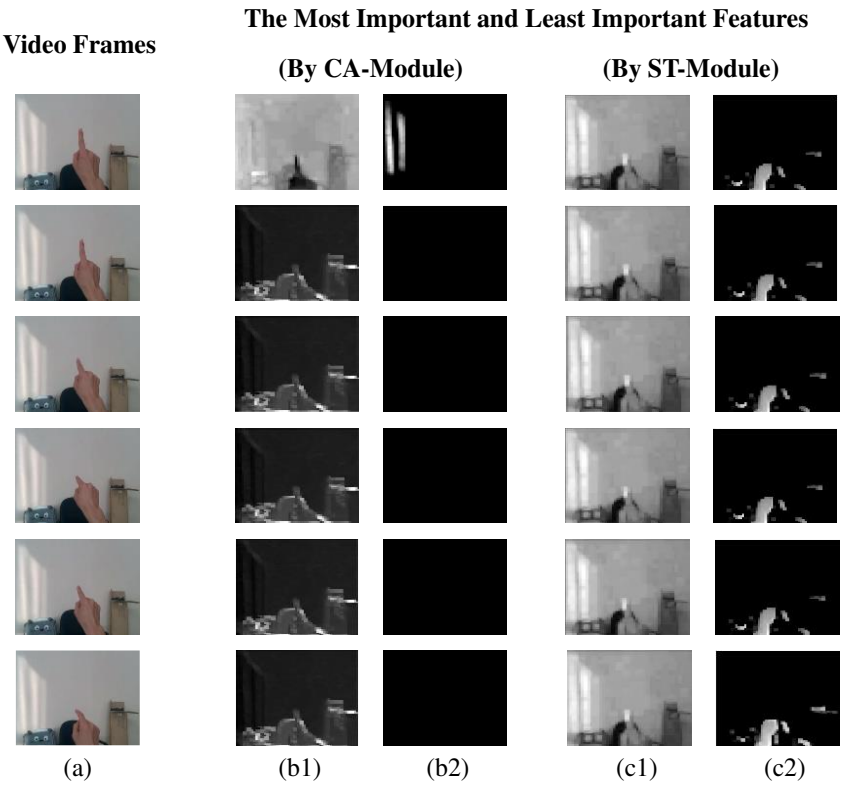

Figure 6. Feature visualization: (a) video frame sequence for the "click with index finger" action; (b1) and (b2) are the highest and lowest weight features determined by the channel attention module; (c1) and (c2) are highest and lowest weighted features determined by the short-term temporal module.

5.2 Training Visualization

In-side (driver) and out-side (road) videos are the most important information in driver intention prediction tasks. In order to effectively model video sequence features, we build an efficient video sequence feature extraction network to extract channel-temporal features of in- and out-side videos respectively and use the EgoGesture[47] dataset to pretrain the constructed network. In the process of building the network, in order to clearly observe the feature extraction operations of the proposed video classification network, we visually displayed the features enhanced and suppressed by the channel attention module and the short-term temporal module through the feature visualization.

As shown in Figure 6, we focus on the most and least important features selected by these two modules. Figure 6(a) shows the video frame sequence for the "click with index finger" action; Figure 6(b1) and (b2) are the highest and lowest weight features determined by the channel attention module; Figure 6(c1) and (c2) are the highest and lowest weighted features determined by the short-term temporal module. It can be seen that the (b1) and (c1) features effectively capture the subject performing the action (“hand”) and the action itself (“click with index finger”). In contrast, the (b2) and (c2) features cannot extract valuable information for action recognition from frame images.

6. Conclusion

In this paper, we propose a driver intention prediction method based on multi-dimensional cross-

modality information interaction. In order to balance the deployment cost of the algorithm while predicting the driver's intention with high accuracy, an efficient video recognition network is designed to extract channel-temporal features of in-side (driver) and out-side (road) videos respectively, in which we design a cross-modality channel-spatial weight mechanism to achieve information interaction between the two feature extraction networks corresponding respectively to the two modalities, and we also introduce a contrastive learning module by which we force the two feature extraction networks to enhance structural knowledge interaction. Then, the obtained representations of in- and out-side videos are fused using a Res-Layer based module to get a preliminary prediction which is then corrected by incorporating the GPS information to obtain a final decision. Besides, we employ a multi-task framework to train the entire network. As demonstrated on the Brain4cars datasets, our driver intention prediction method based on multi-dimensional cross-modality information interaction effectively balances computation and prediction performance, achieving efficient driver behavior prediction.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Reference

1. Bonyani, M.; Rahmanian, M.; Jahangard, S. Predicting Driver Intention Using Deep Neural Network 2021.
2. Rezaei, M.; Klette, R. Look at the Driver, Look at the Road: No Distraction! No Accident! In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; IEEE: Columbus, OH, USA, June 2014; pp. 129–136.
3. Gite, S.; Agrawal, H.; Kotecha, K. Early Anticipation of Driver's Maneuver in Semiautonomous Vehicles Using Deep Learning. *Prog Artif Intell* **2019**, *8*, 293–305, doi:10.1007/s13748-019-00177-z.
4. Yurtsever, E.; Lambert, J.; Carballo, A.; Takeda, K. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access* **2020**, *8*, 58443–58469, doi:10.1109/ACCESS.2020.2983149.
5. Koppula, H.S.; Saxena, A. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 14–29, doi:10.1109/TPAMI.2015.2430335.
6. Gite, S.; Pradhan, B.; Alamri, A.; Kotecha, K. ADMT: Advanced Driver's Movement Tracking System Using Spatio-Temporal Interest Points and Maneuver Anticipation Using Deep Neural Networks. *IEEE Access* **2021**, *9*, 99312–99326, doi:10.1109/ACCESS.2021.3096032.
7. Jain, A.; Koppula, H.S.; Soh, S.; Raghavan, B.; Singh, A.; Saxena, A. Brain4Cars: Car That Knows Before You Do via Sensory-Fusion Deep Learning Architecture 2016.
8. Zhou, D.; Ma, H.; Dong, Y. Driving Maneuvers Prediction Based on Cognition-Driven and Data-Driven Method. In Proceedings of the 2018 IEEE Visual Communications and Image Processing

(VCIP); IEEE: Taichung, Taiwan, December 2018; pp. 1–4.
9. Tonutti, M.; Ruffaldi, E.; Cattaneo, A.; Avizzano, C.A. Robust and Subject-Independent Driving Manoeuvre Anticipation through Domain-Adversarial Recurrent Neural Networks. *Robotics and Autonomous Systems* **2019**, *115*, 162–173, doi:10.1016/j.robot.2019.02.007.
10. Rong, Y.; Akata, Z.; Kasneci, E. Driver Intention Anticipation Based on In-Cabin and Driving Scene Monitoring. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC); IEEE: Rhodes, Greece, September 20 2020; pp. 1–8.
11. Braunagel, C.; Rosenstiel, W.; Kasneci, E. Ready for Take-Over? A New Driver Assistance System for an Automated Classification of Driver Take-Over Readiness. *IEEE Intell. Transport. Syst. Mag.* **2017**, *9*, 10–22, doi:10.1109/MITS.2017.2743165.
12. Jang, Y.-M.; Mallipeddi, R.; Lee, M. Driver's Lane-Change Intent Identification Based on Pupillary Variation. In Proceedings of the 2014 IEEE International Conference on Consumer Electronics (ICCE); IEEE: Las Vegas, NV, USA, January 2014; pp. 197–198.
13. Amsalu, S.B.; Homaifar, A. Driver Behavior Modeling near Intersections Using Hidden Markov Model Based on Genetic Algorithm. In Proceedings of the 2016 IEEE International Conference on Intelligent Transportation Engineering (ICITE); IEEE: Singapore, August 2016; pp. 193–200.
14. Zheng, Y.; Hansen, J.H.L. Lane-Change Detection From Steering Signal Using Spectral Segmentation and Learning-Based Classification. *IEEE Trans. Intell. Veh.* **2017**, *2*, 14–24, doi:10.1109/TIV.2017.2708600.
15. Kim, I.-H.; Bong, J.-H.; Park, J.; Park, S. Prediction of Driver's Intention of Lane Change by Augmenting Sensor Information Using Machine Learning Techniques. *Sensors* **2017**, *17*, 1350, doi:10.3390/s17061350.
16. Chen, H.; Chen, H.; Liu, H.; Feng, X. Spatiotemporal Feature Enhancement Aids the Driving Intention Inference of Intelligent Vehicles. *IJERPH* **2022**, *19*, 11819, doi:10.3390/ijerph191811819.
17. Gite, S.; Agrawal, H. Early Prediction of Driver's Action Using Deep Neural Networks: *International Journal of Information Retrieval Research* **2019**, *9*, 11–27, doi:10.4018/IJIRR.2019040102.
18. Xing, Y.; Hu, Z.; Huang, Z.; Lv, C.; Cao, D.; Velenis, E. Multi-Scale Driver Behaviors Reasoning System for Intelligent Vehicles Based on a Joint Deep Learning Framework. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC); IEEE: Toronto, ON, Canada, October 11 2020; pp. 4410–4415.
19. Bonyani, M.; Rahmanian, M.; Jahangard, S.; Rezaei, M. DIPNet: Driver Intention Prediction for a Safe Takeover Transition in Autonomous Vehicles. *IET Intelligent Trans Sys* **2023**, itr2.12370, doi:10.1049/itr2.12370.
20. Zhou, D.; Liu, H.; Ma, H.; Wang, X.; Zhang, X.; Dong, Y. Driving Behavior Prediction Considering Cognitive Prior and Driving Context. *IEEE Trans. Intell. Transport. Syst.* **2021**, *22*, 2669–2678, doi:10.1109/TITS.2020.2973751.
21. O'Shea, K.; Nash, R. An Introduction to Convolutional Neural Networks 2015.
22. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-Based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. In Proceedings of the Proceedings of the 18th ACM International Conference on Multimodal Interaction; Association for Computing Machinery: New York, NY, USA, October 31 2016; pp. 445–450.
23. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-Temporal LSTM with Trust Gates for 3D Human

Action Recognition. In Proceedings of the Computer Vision – ECCV 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, 2016; pp. 816–833.
24. Stroud, J.; Ross, D.; Sun, C.; Deng, J.; Sukthankar, R. D3D: Distilled 3D Networks for Video Action Recognition.; 2020; pp. 625–634.
25. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features With 3D Convolutional Networks.; 2015; pp. 4489–4497.
26. Lin, J.; Gan, C.; Han, S. TSM: Temporal Shift Module for Efficient Video Understanding.; 2019; pp. 7083–7093.
27. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2014; Vol. 27.
28. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden Two-Stream Convolutional Networks for Action Recognition. In Proceedings of the Computer Vision – ACCV 2018; Jawahar, C.V., Li, H., Mori, G., Schindler, K., Eds.; Springer International Publishing: Cham, 2019; pp. 363–378.
29. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition.; 2016; pp. 1933–1941.
30. Li, Y.; Ji, B.; Shi, X.; Zhang, J.; Kang, B.; Wang, L. TEA: Temporal Excitation and Aggregation for Action Recognition.; 2020; pp. 909–918.
31. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans Pattern Anal Mach Intell* **2020**, *42*, 2011–2023, doi:10.1109/TPAMI.2019.2913372.
32. Liang, Q.; Xiang, S.; Hu, Y.; Coppola, G.; Zhang, D.; Sun, W. PD2SE-Net: Computer-Assisted Plant Disease Diagnosis and Severity Estimation Network. *Computers and Electronics in Agriculture* **2019**, *157*, 518–529, doi:10.1016/j.compag.2019.01.034.
33. Liu, Y.; Ni, K.; Zhang, Y.; Zhou, L.; Zhao, K. Semantic Interleaving Global Channel Attention for Multilabel Remote Sensing Image Classification 2022.
34. T, R.; Valsalan, P.; J, A.; M, J.; S, R.; Latha G, C.P.; T, A. Hyperspectral Image Classification Model Using Squeeze and Excitation Network with Deep Learning. *Comput Intell Neurosci* **2022**, *2022*, 9430779, doi:10.1155/2022/9430779.
35. Perez-Rua, J.-M.; Martinez, B.; Zhu, X.; Toisoul, A.; Escorcia, V.; Xiang, T. Knowing What, Where and When to Look: Efficient Video Action Modeling with Attention 2020.
36. Wang, Z.; She, Q.; Smolic, A. ACTION-Net: Multipath Excitation for Action Recognition.; 2021; pp. 13214–13223.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition.; 2016; pp. 770–778.
38. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks for Action Recognition in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *41*, 2740–2755, doi:10.1109/TPAMI.2018.2868668.
39. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks 2016.
40. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module 2018.
41. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning 2020.
42. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations 2020.

43. Wu, Z.; Xiong, Y.; Yu, S.; Lin, D. Unsupervised Feature Learning via Non-Parametric Instance-Level Discrimination 2018.
44. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.; Li, Z.; Duerig, T. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision 2021.
45. Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.K.; Aggarwal, K.; Som, S.; Wei, F. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts 2022.
46. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation 2014.
47. Zhang, Y.; Cao, C.; Cheng, J.; Lu, H. EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition. *IEEE Transactions on Multimedia* **2018**, *20*, 1038–1050, doi:10.1109/TMM.2018.2808769.