Mixed Spatiotemporal Modeling for Vehicle Behavior Recognition

Ying Zhang[1], Yaochen Li[1⋆], Wei Guo[1], Gaojie Li[1], Shaohan Yang[2], Yuehu Liu[3]
[1]School of Software Engineering, Xi'an Jiaotong University, China
[2]Department of Information and Computing Science, Xi'an Jiaotong-Liverpool University, China
[3]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China

Abstract

Vehicle behavior recognition is a crucial step in perceiving the driving environment for autonomous vehicles. In this paper, we propose a novel mixed spatiotemporal modeling network based on 2D CNN for vehicle behavior recognition. The network is constructed by alternating temporal and spatial modeling to help generate deep and effective spatiotemporal features. For temporal relationships capturing, a Complementary Temporal Extraction (CTE) block is proposed to capture global temporal evolution and motion information in videos, including a Global SpatioTemporal Modeling (GSTM) module, a Motion Excitation (ME) module and a fusion approach designed to merge long and short term features. For spatial information learning, a Channel-enhanced Spatial Extraction (CSE) block is designed, including a modified channel attention module called Spatial Efficient Channel Attention (Spatial-ECA) to learn the relationship between channels. By integrating these modules into the standard residual block with ResNet-50 as the backbone, our method achieves the highest accuracy on vehicle behavior datasets and has a lower model complexity than other 2D CNN-based methods.

1. Introduction

Vehicle behavior recognition is a hot research topic in the communities of intelligent vehicles and computer vision. The detection of driver behaviors can help better understand the traffic scene [18]. The behavior recognition of the vehicle itself and its interaction with the environment can assist the intelligent vehicle to quickly perceive the traffic scene and make better driving decisions. Traffic scene videos from the first perspective are diverse and change rapidly [15]. Therefore, reducing background interference as well as extracting spatial location and temporal motion is important for vehicle behavior recognition.

Recent works [1, 16] using 3D CNN-based frameworks for vehicle behavior recognition are proven to be effective in spatiotemporal modeling. However, the heavy computation and high memory cost are the drawbacks. Multi-modality methods such as [1, 4, 10] utilize optical flows and edge detection images to provide auxiliary spatiotemporal information to the network. However, expensive optical flow computation and complex image processing bring challenges to real-world applications.

Spatiotemporal modeling is crucially important for video action recognition. The MiCT method [30] integrates 2D CNNs with the 3D convolution module to reduce the complexity of spatiotemporal feature fusion and help generate deeper feature maps. The TEA method [11] proposes a *Temporal Excitation and Aggregation* (TEA) block to take both short-term motions and long-term aggregations into consideration. The ACTION-Net method [29] argues that spatiotemporal, channel and motion information are complementary and critical for action recognition.

Inspired by the aforementioned research works, we introduce a small number of 3D convolutional kernels into a 2D CNN-based framework and propose a novel mixed spatiotemporal modeling network, which can effectively capture the spatial information and motion changes of vehicles for vehicle behavior recognition. The main contributions of this paper are summarized as follows:

- A novel mixed spatiotemporal modeling network is proposed. Based on the standard residual block we build two new blocks, the *Complementary Temporal Extraction* (CTE) block and the *Channel-enhanced Spatial Extraction* (CSE) block. Overlaying the two blocks alternately in our network for temporal and spatial modeling can help the network fuse the spatiotemporal features adaptively and learn deeper information.
- The CTE block has both long and short term temporal modeling capabilities, including a *Global SpatioTem-*

*Corresponding author, E-mail: yaochenli@mail.xjtu.edu.cn
This work was supported by the National Key Research and Development Project of New Generation Artificial Intelligence of China under Grant 2018AAA0102504, National Natural Science Foundation of China under Grant No. 61803298, and Key R&D Plan of Shaanxi Province under grant number 2022GY-080.

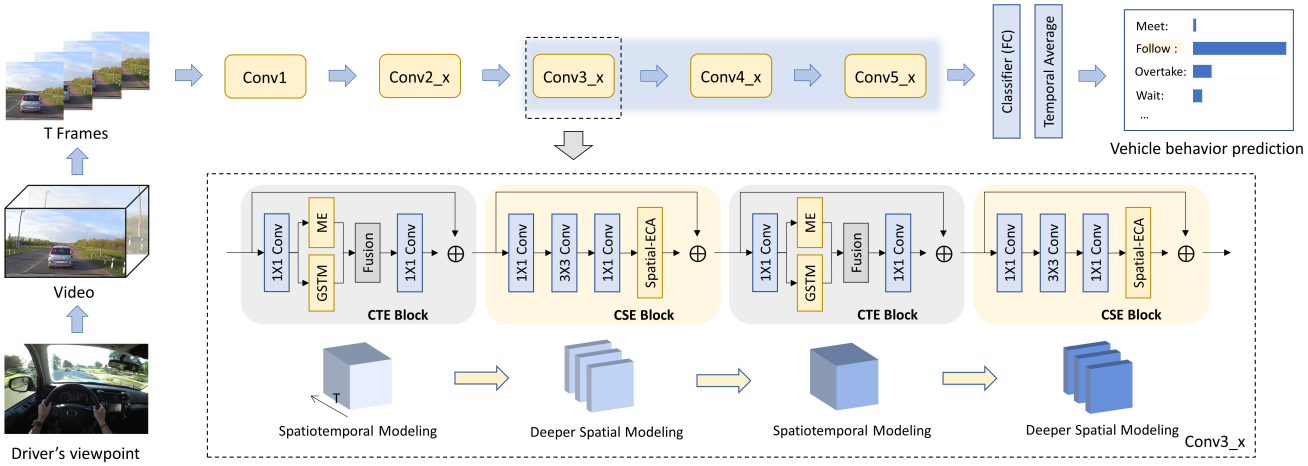

Figure 1. The architecture of our mixed spatiotemporal modeling network. Using 2D ResNet-50 [7] as the backbone, we retain the Conv1 and Conv2_x layers to extract low-level spatial features. Then the CTE block and the CSE block are deployed alternately in the Conv3_x, Conv4_x, and Conv5_x layers to generate deep and effective spatiotemporal feature maps for vehicle behavior recognition.

poral Modeling (GSTM) module, a *Motion Excitation* (ME) module and a fusion approach designed to merge features extracted from the two modules. Meanwhile, the CSE block is designed to conduct spatial modeling with channel information enhanced, including a *Spatial Efficient Channel Attention* (Spatial-ECA) module.

- Our method achieves very high accuracy of vehicle behavior recognition on several traffic datasets. Meanwhile, it shows a significant reduction in FLOPs and parameter numbers than other 2D CNN methods using ResNet-50 [7] as the backbone.

The rest of the paper is organized as follows: The related works are overviewed in Section 2. The proposed method is introduced in Section 3. In Section 4, the experiments and comparisons are conducted, followed by the conclusion in Section 5.

2. Related Works

In this section, we focus on the 2D and 3D CNN-based methods for action recognition and the different scales of temporal modeling.

2.1. 2D CNNs and 3D CNNs

The 2D CNN-based framework is good at spatial feature extraction from a single image but is unable to model the continuous temporal information from image sequences. The C3D method [24] extends the 2D spatial convolution kernels into 3D spatiotemporal convolution kernels, which can effectively extract spatiotemporal features. The FstCN [23] method and the P3D method [17] use a combination of 2D spatial convolution and 1D temporal convolution kernels to replace a 3D convolution kernel. The SlowFast method [3] consists of two pathways with different frame rates that can capture spatial semantics and motion respectively. However, these 3D CNN-based frameworks bring huge memory costs and computation overhead, which prevent the network from generating deeper feature maps for complex tasks and also increase the difficulty of training on large datasets.

The proposal of TSM [12] significantly improves the baseline of 2D CNN-based frameworks for action recognition. The TSM method introduces temporal information into 2D convolution by shifting a part of channels on the temporal dimension, but it has no explicit spatiotemporal modeling. Motivated by the TSM method, many more methods [13, 14, 21, 22] capture temporal information by inserting their designed temporal module into the 2D CNN framework. In particular, the ACTION-Net method [29] designs a plug-and-play module containing spatiotemporal, channel and motion excitation, which can be conveniently and efficiently inserted into 2D CNNs. However, networks constructed in this way lack task-specific design and cannot maximize the extraction and utilization of feature information. Combined models such as ECO [31] and CNN+LSTM [2] use 2D CNN to extract features from each frame independently, and then feed the feature sequence to 3D CNN or RNN for temporal modeling. These methods only consider the temporal relation modeling from high-level spatial features.

2.2. Temporal Modeling

Temporal modeling is one of the priorities for video action recognition. Depending on the scale, temporal modeling can be divided into short-term motion modeling between adjacent frames and long-term temporal modeling

of the whole sequence. The two-stream network [20] uses optical flows as additional inputs to feed the network with pixel-level motion information, while methods [6, 8, 9] extract the motion information of adjacent frames on the feature level. However, these approaches do not consider long-term temporal modeling. Stacking local 3D convolutional kernels repeatedly in deep networks [17, 23, 24] can indirectly model long-term temporal relationships, but also increases the difficulty of optimization. The TEA method [11] proposes motion excitation (ME) and multiple time aggregation (MTA) modules to capture short and long term temporal evolution, but the MTA module is specifically designed for Res2Net [5] and the network lacks deeper spatial modeling. Our method proposes a global spatiotemporal modeling (GSTM) module that introduces a small number of 3D convolution kernels. In order to extract more comprehensive spatiotemporal features at different scales, we design an elegant fusion approach to combine ME and the proposed GSTM module.

3. Our Method

In this section, we describe the mixed spatiotemporal modeling network in detail. Fig. 1 shows the architecture of our network. The *Complementary Temporal Extraction* (CTE) block and the *Channel-enhanced Spatial Extraction* (CSE) block are alternately overlaid in the network to achieve mixed spatiotemporal modeling.

Using the sparse sampling strategy of TSN [27], the first-view vehicle behavior video is sampled to an input image sequence with T frames, and the network predicts the vehicle behavior class end-to-end.

3.1. Complementary Temporal Extraction (CTE)

The CTE block contains a *Global SpatioTemporal Modeling* (GSTM) module to learn long-term temporal evolution, a *Motion Excitation* (ME) module to capture short-term motion information and also a long and short term fusion approach designed to merge the features learned by the two modules.

We integrate the proposed modules into the standard residual block in ResNet-50 to build our CTE block. As illustrated in Fig. 2 (d), the original 3×3 2D convolutional layer in the residual path is replaced by our modules.

3.1.1 Global SpatioTemporal Modeling (GSTM)

The GSTM module is designed for global spatiotemporal modeling. In SlowFast [3], the spatial modeling capability is appropriately weakened and the learning of temporal information is more focused by reducing the channel capacity. Motivated by this, a 1×1 2D convolutional layer $Conv_{reduce}$ is firstly adopted to reduce channels of the input $X \in \mathbb{R}^{N \times T \times C \times H \times W}$, as shown in Fig. 2 (a). It also

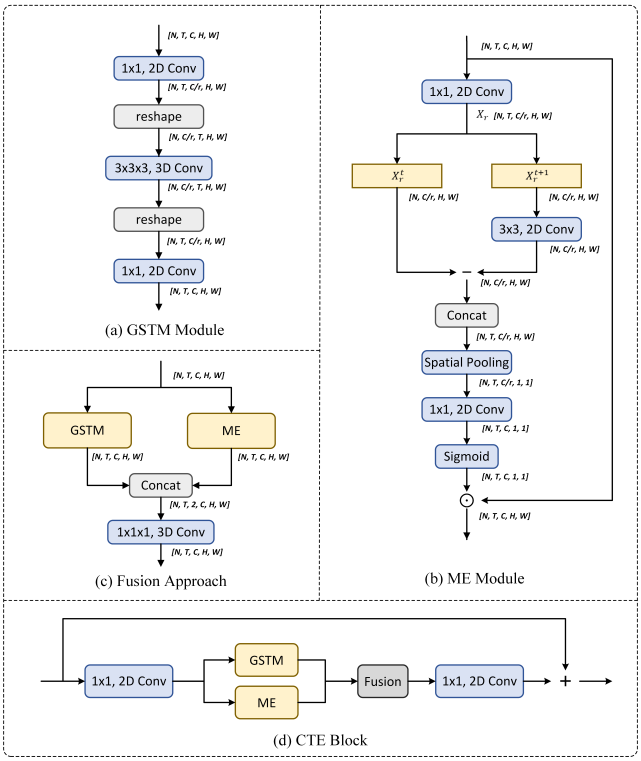

Figure 2. The CTE block.

helps to better control the computational complexity and the number of parameters, striking a balance between efficiency and accuracy.Formally, let X_r denotes the channel-reduced feature and we have:

$$X_r = Conv_{reduce} * X, \quad X_r \in \mathbb{R}^{N \times T \times C/r \times H \times W}, \quad (1)$$

where $*$ indicates the convolution operation, and the reduction ratio is $r = 16$.

The 3D convolution kernels are used to efficiently learn the feature-level global spatiotemporal information from X_r. Then the channel dimension is expanded to the original C to obtain the output M_{glo}, which can be interpreted as:

$$F_{glo} = Conv_{st} * X_r, \quad F_{glo} \in \mathbb{R}^{N \times T \times C/r \times H \times W}, \quad (2)$$

$$M_{glo} = Conv_{expand} * F_{glo}, \quad M_{glo} \in \mathbb{R}^{N \times T \times C \times H \times W}, \quad (3)$$

where $Conv_{st}$ is a $3 \times 3 \times 3$ 3D convolutional layer. $Conv_{expand}$ is a 1×1 2D convolutional layer that is utilized to expand the channel dimension.

3.1.2 Motion Excitation (ME)

The ME module is commonly used in previous works [11, 29] to model motion information of adjacent

frames at the feature level. Unlike these works, we innovatively propose to use ME in parallel with GSTM and also design an efficient fusion method. As illustrated in Fig. 2 (b), we use the difference between two adjacent frames as the feature-level motion representation, which can also reduce the interference of background information. Formally,

$$F_{mo}^{t} = Conv_s * X_r^{t+1} - X_r^t, \quad F_{mo}^t \in \mathbb{R}^{N\times 1\times C/r\times H\times W}, \tag{4}$$

where F_{mo}^t denotes the motion feature between time t and $t+1$, and $Conv_s$ is a 3×3 2D convolutional layer.

We concatenate these motion features according to the temporal dimension and pad the last element with zero i.e., $F_{mo} = [F_{mo}^1, F_{mo}^2, \cdots, F_{mo}^{T-1}, 0], F_{mo} \in \mathbb{R}^{N\times T\times C/r\times H\times W}$. F_{mo} is then processed to $F_{mo}^{'} \in \mathbb{R}^{N\times T\times C/r\times 1\times 1}$ by spatial average pooling. After channel expanding, the motion attention weights A_{mo} can be obtained by using the sigmoid function, which can be represented as:

$$A_{mo} = \sigma(Conv_{expand} * F_{mo}^{'}), \quad A_{mo} \in \mathbb{R}^{N\times T\times C\times 1\times 1}, \tag{5}$$

where σ indicates the sigmoid function. The output of the ME module can be represented as:

$$M_{mo} = X \odot A_{mo}, \quad M_{mo} \in \mathbb{R}^{N\times T\times C\times H\times W} \tag{6}$$

3.1.3 Long and Short Term Feature Fusion

Considering our network is based on the 2D CNN, the feature fusion is designed to be conducted independently in the spatial dimension of each frame in the image sequence.

The fusion method is shown in Fig. 2 (c). The outputs of the GSTM module and the ME module, denoted as $Mglo$ and Mmo, are concatenated, and a $1\times 1\times 1$ 3D convolutional layer $Conv_{concat}$ is used to fuse these features. The final output Y_{st} can be interpreted as:

$$F_{st} = [M_{glo}, M_{mo}], \quad F_{st} \in \mathbb{R}^{N\times T\times 2C\times H\times W}, \tag{7}$$

$$Y_{st} = Conv_{concat} * F_{st}, \quad Y_{st} \in \mathbb{R}^{N\times T\times C\times H\times W} \tag{8}$$

3.2. Channel-Enhanced Spatial Extraction (CSE)

As the channels are repeatedly reduced and expanded in temporal modeling, learning the relationship between channels is necessary. Inspired by the previous work [28] that the Efficient Channel Attention (ECA) module involves only a small number of 1D convolutions while bringing significant performance gains, we make efforts to utilize it in our network. Different from the image classification task, our input $X \in \mathbb{R}^{N\times T\times C\times H\times W}$ has an extra temporal dimension. Therefore, based on whether the module interacts with the temporal dimension, we propose two patterns of the channel attention module: Spatial-ECA and Temporal-ECA. The comparison results are analyzed in Section 4. Since the

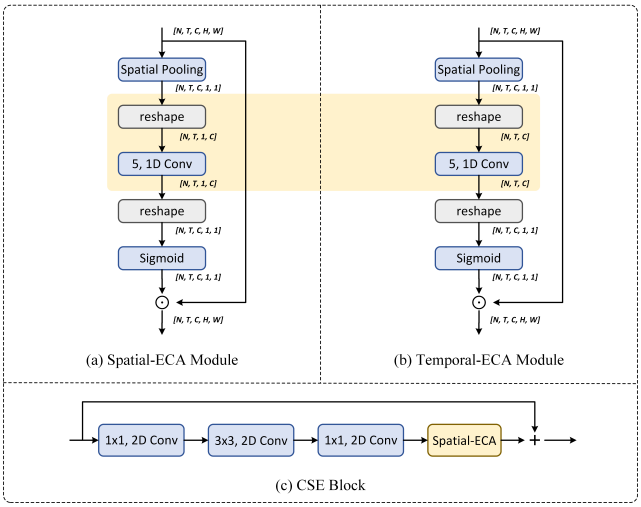

Figure 3. The CSE block.

Spatial-ECA achieves better performance, we form the CSE block by integrating the Spatial-ECA module into the standard residual block. As illustrated in Fig. 3 (c), the channel attention module is inserted after the last 1×1 2D convolutional layer of the residual path.

Specifically, we process the input X to X_p by spatial average pooling as:

$$X_p = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X[:,:,:,i,j], \quad X_p \in \mathbb{R}^{N\times T\times C\times 1\times 1}. \tag{9}$$

Then we reshape X_p to $X_p^{'} \in \mathbb{R}^{N\times T\times 1\times C}$, or reshape X_p directly to $X_p^{''} \in \mathbb{R}^{N\times T\times C}$ to prepare for the following channel attention learning.

After that, a 1D convolutional layer is used to learn the channel attention. Formally, we have:

$$F_{cs} = Conv_{cs} * X_p^{'}, \quad F_{cs} \in \mathbb{R}^{N\times T\times 1\times C}, \tag{10}$$

$$F_{ct} = Conv_{ct} * X_p^{''}, \quad F_{ct} \in \mathbb{R}^{N\times T\times C}, \tag{11}$$

where $Conv_{cs}$ and $Conv_{ct}$ are two different 1D convolutional layers with a kernel size of 5. F_{cs} stands for the feature obtained after the channel attention learning for each frame independently and F_{ct} stands for the feature obtained after the channel attention learning with temporal interaction, corresponding to Fig. 3 (a) and Fig. 3 (b), respectively.

We reshape the tensor $F_c \in \{F_{cs}, F_{ct}\}$ to $F_c^{'} \in \mathbb{R}^{N\times T\times C\times 1\times 1}$, and then use the sigmoid function to obtain the channel attention weights A_c, which can be represented as:

$$A_c = \sigma(F_c^{'}), \quad A_c \in \mathbb{R}^{N\times T\times C\times 1\times 1}, \tag{12}$$

The final output Y_c can be interpreted as:

$$Y_c = X \odot A_c, \quad Y_c \in \mathbb{R}^{N\times T\times C\times H\times W} \tag{13}$$

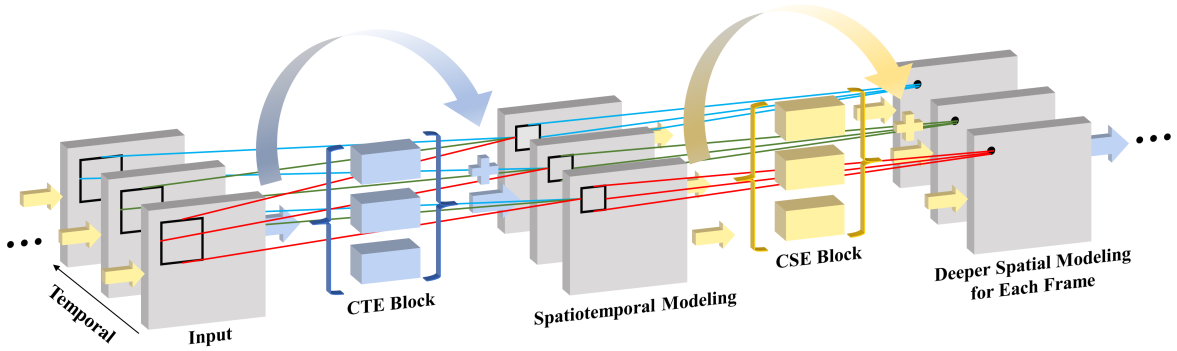

Figure 4. The process of mixed spatiotemporal modeling by using the CTE block and CSE block alternately. On one hand, we conduct spatiotemporal modeling of the deep features. On the other hand, we extract deeper features from the spatial feature maps with temporal information introduced.

Method	Backbone	Pretrain	Frames	FLOPs	ΔFLOPs	Params	ΔParams	BDD100K	
								Top-1	Top-2
TSN [27]	ResNet-50	ImageNet	8	33G	-	23.56M	-	77.86	93.73
TSM [12]	ResNet-50	ImageNet	8	33G	-	23.56M	-	92.62	93.73
ACTION-Net [29]	ResNet-50	ImageNet	8	34.67G	+1.67G(+5.06%)	27.74M	+4.18M(+17.74%)	95.94	98.52
Ours	ResNet-50	ImageNet	8	**27.53G**	**-5.47G(-16.58%)**	**19.28M**	**-4.28M(-18.17%)**	**97.79**	**99.26**

Table 1. **Comparisons with the 2D CNN-based methods.** The model complexity and accuracy of our method are compared with other 2D CNN-based methods on the BDD100K dataset. We re-implement the TSN and TSM methods using the official public code in [12].

3.3. Mixed Spatiotemporal Modeling

Based on the standard residual block [7], we propose the CTE block and the CSE block. The MiCT method [30] verifies that by integrating 2D CNNs to 3D convolution, the network can effectively reduce the complexity of spatiotemporal fusion and generate deeper and more effective feature maps [30]. Rising from the convolution level to the feature extraction level, we insert temporal extraction blocks into the spatial extraction framework for mixed spatiotemporal modeling, which is implemented as alternating the CTE and CSE blocks in 2D ResNet-50.

As shown in Fig. 4, since the advantage of residual connections in blocks is retained, the current CTE block can perform temporal modeling without losing the spatial features learned in the previous CSE block. Mixed modeling allows the network to fuse spatiotemporal features adaptively and accelerates the optimization of the whole network.

4. Experiments

In this section, we first introduce the evaluation datasets and the implementation details. We then evaluate the performance of our method by comparing with the 2D CNN-based and the state-of-the-art approaches. Also, we conduct an intuitive ablation experiment for all modules used in our method including GSTM, ME and Spatial-ECA. Moreover, experiments are conducted to evaluate the effectiveness of the two different attention modules designed in the CSE block and the feature fusion approaches in the CTE block. Finally, the visualization results are presented.

4.1. Datasets

The experiments are conducted on the HDD+ dataset and the BDD100K dataset annotated in the previous work [10]. The HDD+ dataset includes 6 categories of common vehicle behaviors: turning left, turning right, overtaking, being overtaken, following and meeting, representing the behaviors of the vehicle itself and its interaction with surrounding vehicles in first-view videos. The BDD100K dataset covers more diverse class divisions under complex scenes such as rain and night which contain 8 categories including overtaking, being overtaken (from the left or right), following, meeting, turning left, turning right and waiting.

4.2. Implementation Details

We follow the sparse sampling strategy of TSN [27] to process the video into the input of the network. Specifically, we first split the input video into T segments of equal length, and then randomly select one frame from each segment to obtain a sparse sequence with T frames. The shorter side of these frames is fixed to 256 and corner cropping is utilized for data augmentation. Finally, each frame is resized to 224×224 for training. The input of the model has

Method	Backbone	Pretrain	Frames	FLOPs	Params	BDD100K			HDD+		
						Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
C3D [24]	-	-	16	38.55G	78.38M	81.92	86.35	91.14	87.25	90.49	95.41
R3D [25]	3D ResNet-18	-	16	40.89G	33.18M	87.82	92.25	95.94	93.75	94.10	98.36
SELayer-3DCNN [10]	-	-	16	**10.77G**	**3.85M**	90.04	95.94	97.05	92.44	93.44	98.69
TSM [12]	ResNet-50	ImageNet	8	33G	23.56M	92.62	98.52	99.26	96.39	98.63	99.48
TEA [11]	ResNet-50	ImageNet	8	34.74G	24.12M	92.99	97.05	99.26	96.56	98.97	99.31
TDN [26]	ResNet-50	ImageNet	8	36G	24.07M	94.83	98.16	99.26	96.22	98.80	99.66
ACITON-Net [29]	ResNet-50	ImageNet	8	34.67G	27.74M	95.94	98.52	99.63	97.77	98.97	100.00
Ours	ResNet-50	ImageNet	8	27.53G	19.28M	**97.79**	**99.26**	**99.63**	**97.94**	**99.66**	**100.00**

Table 2. **Comparisons with the state-of-the-arts.** All the comparison methods are re-implemented using their official public codes with the same data augmentation as our method.

GSTM	ME	Spatial-ECA	BDD100K		HDD+	
			Top-1	Top-2	Top-1	Top-2
			77.86	93.73	84.54	97.42
✓			94.46	98.52	97.08	99.14
✓	✓		94.83	98.89	97.08	99.48
✓		✓	95.20	99.26	97.42	99.66
✓	✓	✓	**97.79**	**99.26**	**97.94**	**99.66**

Table 3. **Ablation studies of the proposed modules.** The GSTM module and the ME module are in the CTE block, and the Spatial-ECA module is in the CSE block.

Channel Attention	BDD100K		HDD+	
	Top-1	Top-2	Top-1	Top-2
Without ECA	94.83	98.89	97.08	99.48
Temporal-ECA	95.20	99.26	97.08	99.31
Spatial-ECA	**97.79**	**99.26**	**97.94**	**99.66**

Table 4. **Ablation studies on different channel attention modules.** Temporal-ECA performs channel attention learning interacting with the temporal dimension, while Spatial-ECA performs inter-frame independent channel attention learning.

Feature Fusion	BDD100K		HDD+	
	Top-1	Top-2	Top-1	Top-2
Add(GSTM, ME)	93.73	98.15	96.91	99.48
Ours(GSTM, ME)	**97.79**	**99.26**	**97.94**	**99.66**

Table 5. **Ablation studies on different fusion approaches.** In this table, Add(·) refers to directly adding the outputs of the two modules GSTM and ME, while Ours(·) refers to our proposed long and short term fusion approach.

the size of $N \times T \times 3 \times 224 \times 224$, where N refers to the batch size and T to the number of segments.

We use Adam optimizer to train the network and set $N = 8$, $T = 8$. The weights of the network are initialized with the pre-trained Resnet-50 on ImageNet. We start with a learning rate of 0.0001 and reduce it by a factor of 10 at 25, 35 epochs with 50 epochs in total.

4.3. Comparisons with 2D CNNs

We compare our method with three representative 2D CNN-based methods TSN, TSM and ACTION-Net. As illustrated in Table 1, our method has the highest performance and lowest complexity, which is very cost-effective. The TSN method does not contain any temporal modeling component and hence results in lower accuracy. The TSM method introduces implicit temporal modeling with zero parameter increase in TSN and achieves a significant improvement. The ACTION-Net method adds their designed module to TSN, which improves the performance but also brings some increase in parameter number. In contrast, our method does not simply add new modules but replaces the blocks in TSN, achieving high performance while reducing the model complexity. Specifically, our method reduces 16.58% in FLOPs and 18.17% in parameter number compared to TSN and TSM. Compared to ACTION-Net, our method improves 1.85% in Top-1 accuracy and 0.74% in Top-2 accuracy on the BDD100K dataset.

4.4. Comparisons with the state-of-the-arts

We compare our method with the state-of-the-art approaches on two datasets. As shown in Table 2, our method achieves the highest accuracy on both the BDD100K and HDD+ datasets, and has significant advantages on computational complexity and parameter number, with an impressive performance on the vehicle behavior recognition task. Although the SELayer-3DCNN method has a more lightweight network, it introduces additional optical flow and edge images to form the multi-stream input along with RGB images, which is time-consuming in image preprocessing. Our method only takes the RGB images as input and obtains higher accuracy. Compared with other single-

stream input approaches, either 3D or 2D CNN-based, our method has the lowest model complexity and the highest accuracy for vehicle behavior recognition.

4.5. Ablation Study

Contribution of the proposed modules. To further analyze the specific roles and contributions of the GSTM, ME and Spatial-ECA modules proposed in our method, we conduct a comprehensive ablation experiment on the BDD100K and HDD+ datasets. As shown in Table 3, the model with ResNet-50 as backbone has a very significant performance improvement after adding the GSTM module for global spatiotemporal modeling of image sequences, with Top-1 and Top-2 accuracy on the BDD100K dataset rising by 16.60% and 4.79%, and Top-1 and Top-2 accuracy on the HDD+ dataset rising by 13.40% and 1.74%. This proves that the GSTM module has a good spatiotemporal modeling capability. After adding the ME module for short-term motion information supplementation and the spatial-ECA module for channel importance learning, the Top-1 accuracy on the BDD100K and HDD+ datasets obtains an additional improvement by 3.33% and 0.86%, respectively.

Comparison of the different channel attention modules. We conduct an experiment to compare the performance of the two different channel attention modules, Spatial-ECA and Temporal-ECA. As illustrated in Table 4, the Spatial-ECA achieves better results on both datasets. We suggest that temporal information has been merged into the spatial features of each frame in spatiotemporal modeling. Therefore, interacting with the temporal dimension while learning the relationship between channels in spatial modeling may lead to interference instead.

Effectiveness of the feature fusion approach. Finally, we conduct an ablation experiment to verify the effectiveness of the long and short term fusion approach in the CTE block. Our fusion approach is designed to merge the spatiotemporal features of different terms extracted from the GSTM and ME modules. As can be seen from Table 5, compared to directly adding the outputs of the GSTM and ME modules to fuse features, our proposed fusion approach is able to maximize the complementary information learned by the two modules.

4.6. Visualization Results

As shown in Fig. 5, we visualize the significant features of different vehicle behaviors extracted by our method. For the interaction behaviors including overtaking, being overtaken (from left or right), following and meeting, our method can effectively focus on spatial location and motion changes of the interactive vehicle. For the driving behaviors of the vehicle itself, certain static traffic elements in the driving context will be seen as references. For behaviors such as turning left, turning right and waiting, the visualization results are displayed as the right shift, left shift and no shift of the reference. It is interesting to note that the re-

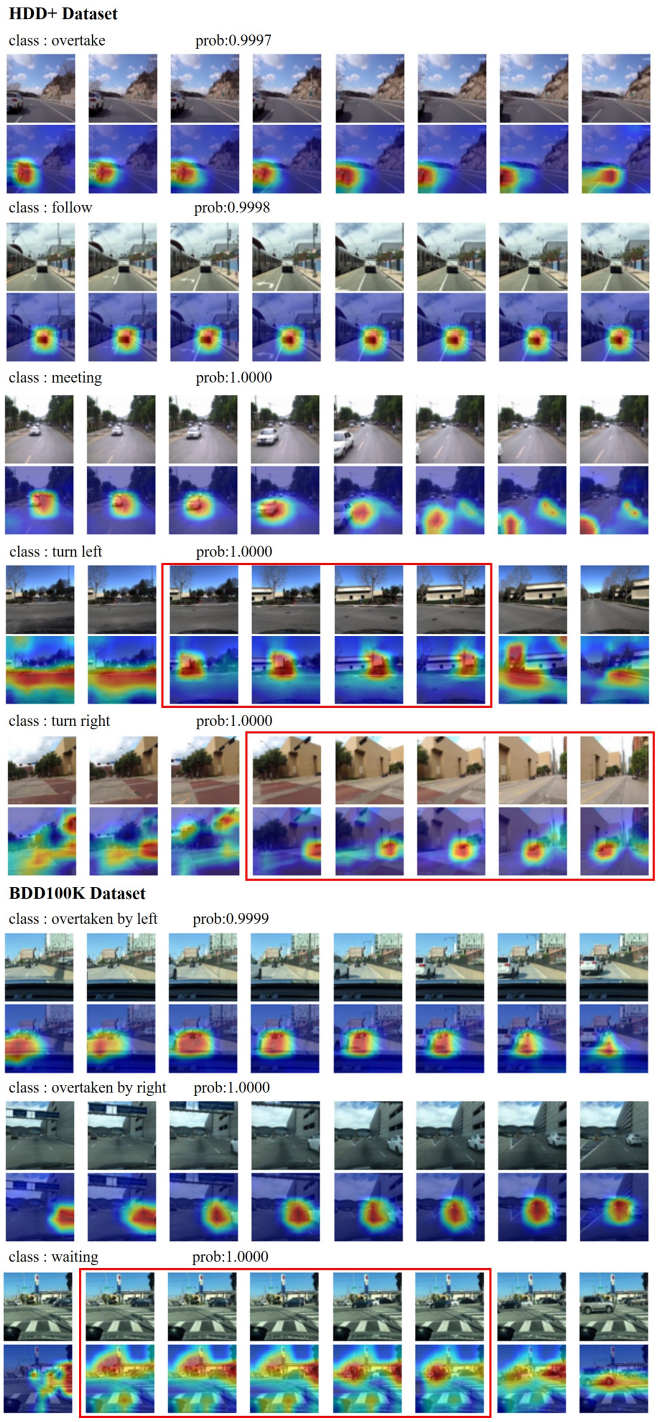

Figure 5. Visualization for significant features on the HDD+ and BDD100K datasets extracted by our method using Grad-CAM [19]. The odd rows indicate the raw images, and the even rows indicate the results of feature visualization.

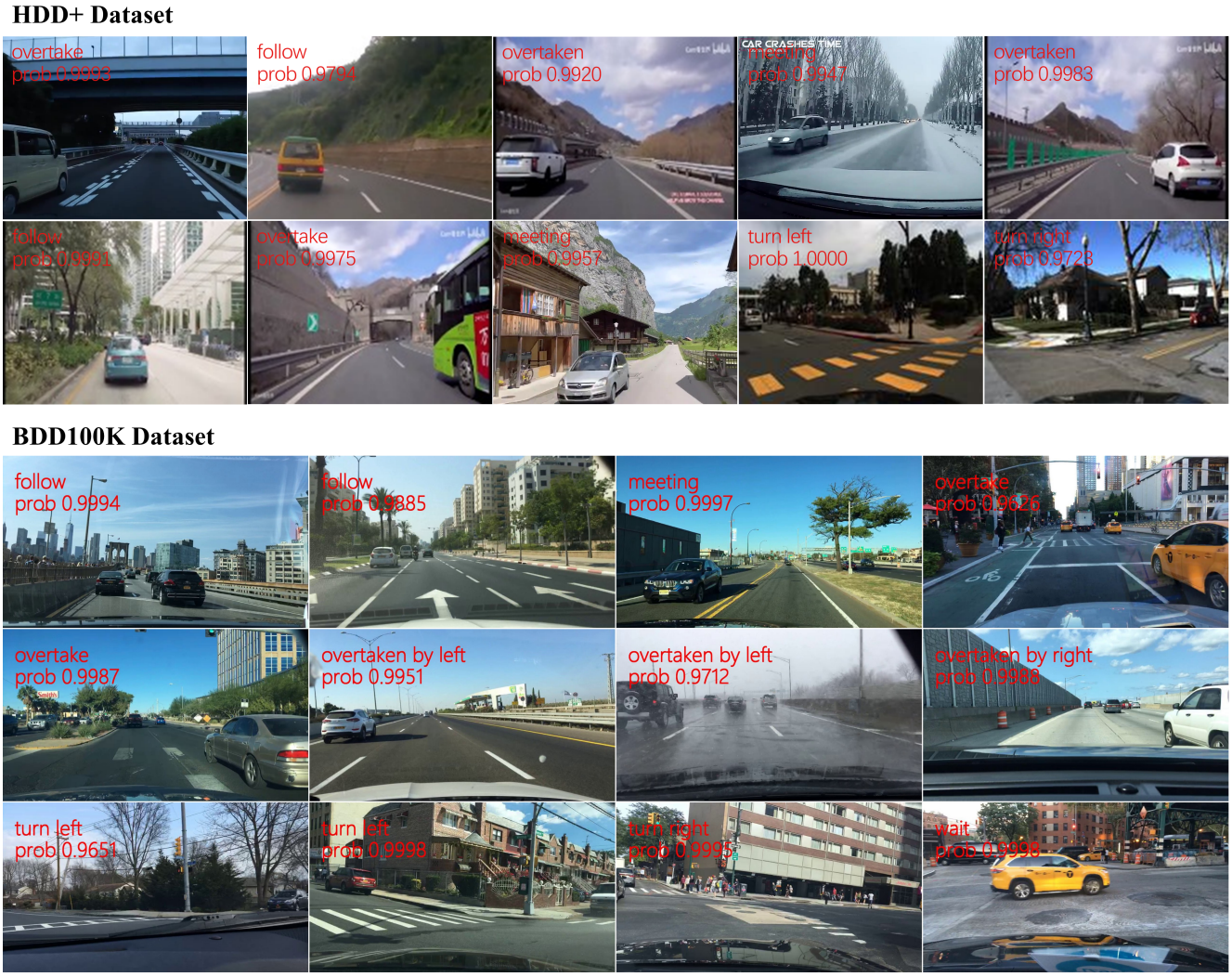

Figure 6. The visualization results of vehicle behavior recognition on the HDD+ and BDD100K datasets using the proposed method.

sults, especially in the categories of following, meeting and waiting, demonstrate that vehicle behavior recognition has positive effects on tasks such as vehicle object tracking, lane line detection and semantic segmentation in traffic scenes.

The final visualization results of the proposed vehicle behavior recognition method on the HDD+ and BDD100K datasets are shown in Fig. 6.

5. Conclusion

In this paper, we propose a new mixed spatiotemporal modeling network for vehicle behavior recognition. By mixed spatiotemporal modeling, the network generates deep and effective spatiotemporal features with smaller fusion complexity to help classify vehicle behavior. Specifically, a CTE block is proposed to capture the long-term spatiotemporal relationship and short-term motion information. Also, a CSE block is designed for spatial modeling, including a Spatial-ECA module to learn the relationship between channels.

More vehicle behaviors will be investigated in the future. And with the foundation of our vehicle behavior recognition work, we will explore vehicle behavior prediction and multi-task learning to assist autonomous vehicles achieve faster and more comprehensive detection and understanding of the traffic scene.

References

[1] Mahdi Biparva, David Fernández-Llorca, Rubén Izquierdo Gonzalo, and John K. Tsotsos. Video action recognition for lane-change classification and prediction of surrounding vehicles. *IEEE Transactions on Intelligent Vehicles*, 7(3):569–578, 2022. 1

[2] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015. 2

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019. 2, 3

[4] David Fernández-Llorca, Mahdi Biparva, Rubén Izquierdo-Gonzalo, and John K. Tsotsos. Two-stream networks for lane-change prediction of surrounding vehicles. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, 2020. 1

[5] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2021. 3

[6] Brennan Gebotys, Alexander Wong, and David A Clausi. M2a: Motion aware attention for accurate video action recognition. In *2022 19th Conference on Robots and Vision (CRV)*, pages 83–89. IEEE, 2022. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5

[8] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2000–2009, 2019. 3

[9] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018. 3

[10] Yaochen Li, Haochuan Hou, Zikun Dong, Yujie Zang, Ying Zhang, and Yonghong Song. Spatiotemporal analysis of static and dynamic traffic elements from road scenes. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–15, 2022. 1, 5, 6

[11] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 906–915, 2020. 1, 3, 6

[12] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092, 2019. 2, 5, 6

[13] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11669–11676, 2020. 2

[14] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13708–13718, 2021. 2

[15] Athma Narayanan, Isht Dwivedi, and Behzad Dariush. Dynamic traffic scene classification with space-time coherence. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5629–5635, 2019. 1

[16] Alam Noor, Bilel Benjdira, Adel Ammar, and Anis Koubaa. Driftnet: Aggressive driving behaviour detection using 3d convolutional neural networks. In *2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pages 214–219, 2020. 1

[17] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5534–5542, 2017. 2, 3

[18] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018. 1

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7

[20] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 568–576, 2014. 3

[21] Haisheng Su, Kunchang Li, Jinyuan Feng, Dongliang Wang, Weihao Gan, Wei Wu, and Yu Qiao. Tsi: Temporal saliency integration for video action recognition. *arXiv preprint arXiv:2106.01088*, 2021. 2

[22] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1102–1111, 2020. 2

[23] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E. Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4597–4605, 2015. 2, 3

[24] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 2, 3, 6

[25] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6

[26] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1895–1904, 2021. 6

[27] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 3, 5

[28] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539, 2020. 4

[29] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13209–13218, 2021. 1, 2, 3, 5, 6

[30] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Mict: Mixed 3d/2d convolutional tube for human action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 449–458, 2018. 1, 5

[31] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018. 2