TEM^3-Learning: Time-Efficient Multimodal Multi-Task Learning Network for Assistive Driving

Wenzhuo Liu[1], Yicheng Qiao[2], Zhen Wang[1], Qiannan Guo[2], Zilong Chen[2], Meihua Zhou[2], Xinran Li[3], Letian Wang[4], Zhiwei Li[5], Huaping Liu[2], Wenshuo Wang[1,*]

***Abstract*— While multi-task learning (MTL) improves driver assistance systems by exploring inter-task correlations through shared representations, existing methods suffer from two critical limitations: (1) single-modality constraints that restrict comprehensive scene understanding, and (2) inefficient architectures hindering real-time deployment. This paper proposes a Time-Efficient Multimodal Multi-task Learning, called TEM^3-Learning, framework that jointly optimizes driver emotion recognition, driver behavior classification, traffic context understanding, and vehicle behavior prediction through a novel two-stage architecture. The first component, the mamba-based multi-view temporal-spatial feature extraction subnetwork (MTS-Mamba), introduces a forward-backward temporal scanning mechanism and global-local spatial attention, achieving low-cost temporal-spatial feature extraction from multi-view sequential images. The second component, the MTL-based gated multimodal feature integrator (MGMI), employs task-specific multi-gating modules to adaptively emphasize the most relevant modality features for each recognition task to alleviate the negative transfer problem in MTL. Evaluation on the AIDE dataset, our proposed model attains state-of-the-art accuracy across all four tasks with less than 6M parameters and achieves 142.32 FPS inference speed. Rigorous ablation studies confirm the effectiveness of the proposed method and each module's independent contributions.**

I. INTRODUCTION

Advanced Driver Assistance Systems (ADAS) improve driving safety by continuously monitoring the driver's state and traffic environment [1], [2]. However, existing research is mostly limited to the solution of single tasks, such as driver emotion/behavior, traffic environment recognition, without addressing the inherent interdependencies between these tasks [3]. Studies show a significant coupling relationship between the driver's state and the traffic environment [4], [5]. For example, drivers frequently adjust their behavior, such as changing lanes based on surrounding traffic conditions, while traffic congestion can induce driver anxiety. Integrating driver state and traffic environment recognition into a unified multi-task learning (MTL) framework can provide a more holistic understanding of driving scenarios and enhance the safety performance of ADAS [4].

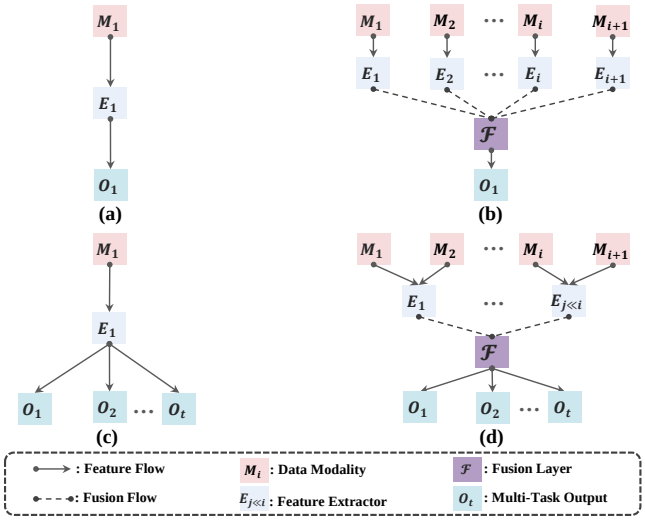

Fig. 1. Comparison diagram of four mainstream algorithm frameworks: (a) single-modal single-task, (b) multimodal single-task, (c) single-modal multi-task, and (d) the multimodal multi-task algorithm proposed in this paper. Through comparative analysis, it can be observed that as the number of tasks and modal dimensions increases, the complexity of the model structure rises significantly. Therefore, ensuring computational efficiency while improving model accuracy becomes a key challenge in multimodal multitask learning.

Compared to single-task learning, MTL reduces overfitting by sharing features across tasks, thereby improving the performance of each individual task [6]. In the context of ADAS, existing MTL models generally focus on related sub-tasks. For example, Wu proposed a model that jointly learns multiple traffic environment-related tasks, enabling simultaneous recognition of lane markings, drivable areas, and object detection in driving scenes [7]. Similarly, Xing et al. focus on driver-related MTL to recognize the driver's emotions and behaviors [8]. While these approaches have yielded performance improvements, they often struggle with negative transfer, where task performance deteriorates due to task-related conflicts or differences[9].

MTL models tend to have more complex structures compared to single-task learning, as they must simultaneously handle multiple tasks [7] (see Fig. 1 (a) (c)). Many existing

*denotes the corresponding author.

[1]Wenzhuo Liu, Zhen Wang, and Wenshuo Wang are with the Faculty of Marine Science and Technology, Beijing Institute of Technology, Zhuhai, China, 519088 (e-mail:wzliu@bit.edu.cn; 3220245402@bit.edu.cn; ws.wang@bit.edu.cn).

[2]Yicheng Qiao, Qiannan Guo, Zilong Chen, Meihua Zhou, and Huaping Liu are with the State Key Laboratory of Intelligent Technology and Systems and Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China (e-mail: yichengqiao21@gmail.com; guoqiannan1203@163.com; chenzl22@mails.tsinghua.edu.cn; mhzhou0412@gmail.com; hpliu@tsinghua.edu.cn).

[3]Li Xinran is with the Graduate School of Arts and Sciences, Yale University, New Haven, CT, USA (e-mail: xinranll668@gmail.com).

[4]Letian Wang is with the University of Toronto, Toronto, M5S2E8, Canada (e-mail: lt.wang@mail.utoronto.ca).

[5]Zhiwei Li is with Beijing University of Chemical Technology, Beijing, 100029, China (e-mail: 2022500066@buct.edu.cn).

MTL methods rely on inputs from a single modality such as driving scene images [10]. However, optimal task performance requires complementary multimodal inputs[11], [12]. For example, combining driver images with eye movement data can improve emotion recognition [13], while fusing image and point cloud data can enhance traffic object recognition [14]. Unfortunately, many multimodal input algorithms still rely on independent feature extraction methods [15], [16], [17], which significantly increase model parameters, reduce inference speed, and compromise practical scalability (see Fig. 1 (b)). Thus, a critical challenge is how to effectively utilize multimodal information for MTL while optimizing both individual task performance and task-level synergy, all while maintaining high inference speed and a low parameter count.

To address these challenges, we propose a Time-efficient multimodal multi-task learning network that simultaneously performs four tasks — driver emotion recognition (DER), driver behavior recognition (DBR), traffic context recognition (TCR), and vehicle behavior recognition (VBR) — using multimodal data (Fig. 1 (d)). This is achieved through two key components: the Mamba-based multi-view temporal-spatial feature extraction subnetwork (MTS-Mamba) and the multi-task learning-based gated multimodal feature integrator (MGMI). MTS-Mamba introduces a forward-backward temporal scanning mechanism and a global-local spatial feature extraction strategy, enabling efficient extraction of temporal-spatial features from multi-view sequential images at low computational cost. This approach provides richer and more robust feature representations for subsequent multi-task recognition. MGMI, inspired by previous work [18], incorporates a multi-gating mechanism that dynamically adjusts the weights of modality features based on task-specific attention. This selective reinforcement of important features helps alleviate negative transfer. We validated the effectiveness of our method using the publicly available AIDE dataset. Experimental results show that our model outperforms previous methods on all four tasks, with fewer than 6M parameters and exceptionally fast inference speed, demonstrating its efficiency and practicality. Our contributions include:

- A time-efficient multimodal MTL framework that provides a new paradigm for multimodal MTL in ADAS.
- The MTS-Mamba subnetwork, which effectively extracts temporal-spatial features from multi-view sequential images across multiple dimensions.
- The MGMI mechanism, which adaptively adjusts attention to different modalities for each task, mitigating negative transfer and enhancing task-specific feature extraction.

II. Related Work

A. Multimodal Learning

ADAS systems that rely solely on single-modality data often struggle to handle the complex and dynamic challenges encountered during driving [19], [20], [13], [21], [22]. Each modality offers unique advantages and limitations, and leveraging multiple modalities can provide a more comprehensive understanding of the driving environment. For example, combining multi-view driving scene images with LiDAR data enhances environmental perception [23]. Similarly, integrating driver images and joint information can effectively capture facial expressions and behavioral states [4]. As a result, many studies have increasingly turned to multi-modal data to enhance task accuracy. For instance, Zhou et al. used front-view driving scene images, driver images, and vehicle speed data for driver behavior recognition [24], while Liu et al. combined camera images and LiDAR data for 3D object detection[25].

However, multimodal learning in ADAS faces two key challenges. First, many existing models rely on independent feature extraction branches for each modality, even when processing multi-view images [16], [11], [24]. This approach significantly increases the model's parameter count and overlooks potential intermodal interactions, leading to inefficient information use. Second, multimodal fusion methods are often limited to combining only a few modality features, restricting their generalization ability and scalability.

B. Multi-task Learning

MTL improves model performance on each task by sharing features across tasks [6]. The two predominant strategies for parameter sharing are hard and soft parameter sharing. In hard parameter sharing, most parameters are shared across tasks, with task-specific differentiation occurring only in the final layers [26], [27], [28]. For example, Wu et al. adopted this structure to simultaneously achieve traffic object detection, drivable area segmentation, and lane detection [7]. While simple and efficient, this approach is prone to negative transfer when tasks exhibit substantial differences, limiting its generalizability. To mitigate this, soft parameter sharing has been proposed, where each task retains independent parameters but leverages shared features, thus preserving task independence and reducing task conflicts [29]. For instance, Choi et al. designed a task-adaptive attention generator to enable independent parameters for tasks like monocular 3D object detection, semantic segmentation, and dense depth estimation [10].

Although soft parameter sharing alleviates task conflicts, two critical challenges remain. First, introducing task-specific parameters increases the model's flexibility but significantly raises the parameter count, which can hinder real-time performance — a major concern in applications like ADAS, where real-time processing is crucial. Second, existing MTL models in ADAS are predominantly designed for single-modality inputs (e.g., driving scene images), limiting their applicability to multimodal scenarios and constraining the full potential of MTL-based solutions.

III. Methodology

This section presents the overall structure and key modules of the proposed TEM^3-Learning network. We first describe

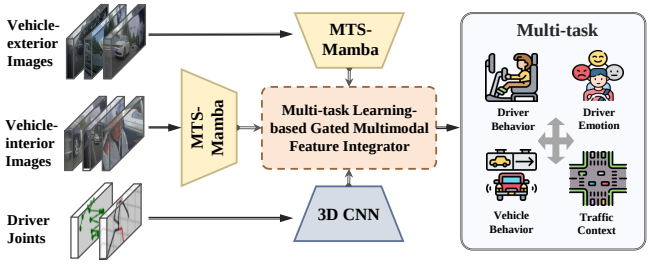

Fig. 2. The overall pipeline of TEM^3-Learning. MTS-Mamba and 3D CNN are used to extract multimodal features from vehicle-exterior images, vehicle-interior images, and driver joints, respectively. Subsequently, the multi-task learning-based gated multimodal feature integrator (MGMI) adaptively fuses these features, ultimately enabling multitask recognition.

the overall architecture of the network and then detail two core modules: MTS-Mamba and MGMI.

A. Network Overview

To fully leverage the synergies between multimodal and multi-view data while balancing accuracy and efficiency in a MTL context, our proposed TEM^3-Learning network (Fig. 2) consists of three core components: the multi-branch feature extraction layer, MGMI, and the multi-task recognition layer.

The first component, the multi-branch feature extraction layer, processes multimodal data through two modules: MTS-Mamba and 3D CNN. MTS-Mamba extracts deep temporal-spatial features from vehicle-exterior images (front, left, and right views) and vehicle-interior images (driver's facial and body images). The 3D CNN module focuses on extracting high-level semantic features from the driver's posture and gestures. These extracted features are then integrated by MGMI, which first uses a self-attention mechanism to obtain task-shared features, followed by a multi-gating mechanism to adaptively weight features based on task-specific attention, thus improving feature extraction and alliating conflicts between tasks.

During training, we use a cross-entropy loss function that integrates individual task losses to optimize the overall model performance. The total loss L_{total} is computed as:

$$L_{\text{total}} = \sum_{r=1}^{m} \text{CrossEntropy}\Big(\hat{y}_r, y_r\Big), \tag{1}$$

where $\hat{y}_r$ denotes the recognition results for task r, and y_r denotes the corresponding ground truth. The number of tasks m is set to 4, corresponding to the four recognition tasks: DER, DBR, TCR and VBR.

B. MTS-Mamba

Multi-view sequential images from both vehicle interiors and exteriors exhibit strong temporal-spatial correlations. For instance, sequential vehicle-interior images effectively capture driver behavior (e.g., looking around, making calls), while multi-view vehicle-exterior images provide a comprehensive understanding of surrounding environments (e.g., pedestrians, vehicles, obstacles). Leveraging these temporal-spatial features is crucial for enhancing ADAS performance in environmental perception and behavior recognition. However, in real-world driving scenarios — especially in complex scenarios like congested intersections where dynamic targets and complex movement patterns are prevalent — modeling these features becomes increasingly challenging. The difficulty arises from the drastic environmental changes and the need to balance modeling quality with real-time performance. CNNs offer good real-time performance but are limited by their local receptive field, which hampers the capture of global temporal-spatial features and, consequently, task recognition accuracy. On the other hand, attention-based methods, while capable of global modeling, suffer from high computational complexity, making them unsuitable for processing large volumes of multi-view sequential data in real-time ADAS applications. To address these challenges, we propose MTS-Mamba, which combines a dual-path temporal-spatial feature extraction structure with a State Space Model (SSM) based on Mamba [30]. This approach efficiently captures multi-view sequential image features while maintaining low computational cost, ensuring a balance between recognition accuracy and real-time performance.

Before MTS-Mamba, we need to process multi-view sequential images. First, the sequential images of the j-th view are concatenated along the channel dimension, forming a sequence $\mathbf{I}_j \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the channel count, height, and width, respectively. Each $\mathbf{I}_j$ is then processed through deep convolution and adaptive average pooling to extract initial features of identical dimensions. These features are integrated into a feature map $\mathbf{F}_f$, which is fed into the MTS-Mamba module for further processing.

For the input feature map $\mathbf{F}_f$, we first apply 1D convolution with GELU activation to enhance feature representation within the dual-path temporal-spatial feature extraction structure,i.e., global and local layers. The local and global layers employ forward and backward scanning, respectively, to extract bidirectional temporal features. Specifically, given that $\mathbf{F}_f$ spans 16 consecutive frames, the local layer perform a forward scan where each channel of $\mathbf{F}_f$ is linearly transformed by the state parameters $\mathbf{B}\cdot\mathbf{C}^\top$ of the state-space model, capturing forward temporal dependencies. In the global layer, the spatial order of $\mathbf{F}_f$ is reversed during the backward scanning to capture temporal interdependencies from both directions. Then, it is linearly transformed with $\mathbf{B}\cdot\mathbf{C}^\top$. During training, the shared state parameters $\mathbf{B}$ and $\mathbf{C}$ are updated via backpropagation to retain bidirectional temporal information. The SSM calculates the temporal feature weight $\mathcal{W}_{\text{ssm}}$ as:

$$\begin{cases} \mathbf{B}^{(s+1)} = \mathbf{B}^{(s)} - \eta\cdot\dfrac{\partial\mathcal{L}}{\partial\mathbf{B}^{(s)}}, \\ \mathbf{C}^{(s+1)} = \mathbf{C}^{(s)} - \eta\cdot\dfrac{\partial\mathcal{L}}{\partial\mathbf{C}^{(s)}}, \end{cases} \tag{2}$$

$$\mathcal{W}_{\text{ssm}} = \sigma\left(\mathbf{A}\cdot\mathbf{d}_{\text{state}} + \mathbf{B}\cdot\mathbf{C}^{\text{T}}\cdot\mathbf{d}_{\text{dim}} + \mathbf{D}\right), \tag{3}$$

where $\mathbf{B}^{(t+1)}$ denotes the updated shared weight parameter, η is the learning rate, $\sigma(\cdot)$ is the sigmoid function, $\mathbf{A}\in$

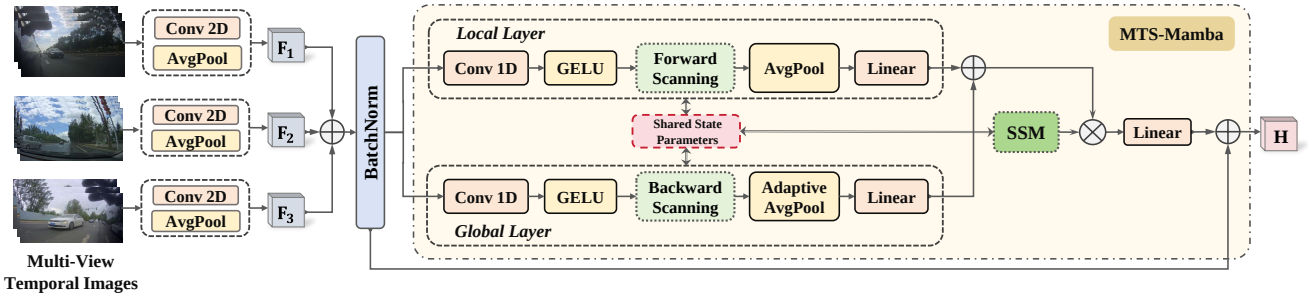

Fig. 3. The structure diagram of MTS-Mamba, which includes the forward and backward scanning mechanisms, as well as the global-local spatial feature extraction strategy.

$\mathbb{R}^{d\times n}$ is the state transition matrix, $\mathbf{d}_{\text{state}}$ and $\mathbf{d}_{\text{dim}}$ are the dimensions of states and input features, and $\mathbf{D} \in \mathbb{R}^d$ is the bias matrix.

After capturing the bidirectional temporal features, the local layer uses 3×3 average pooling and linear projection to extract local spatial features $\mathbf{F}_l$, while the global layer uses adaptive average pooling and linear projection to capture global spatial features $\mathbf{F}_g$. This captures detailed features and global context from spatial information at different scales, providing more comprehensive and rich spatial feature representations. We then merge the multi-scale spatial features $\mathbf{F}_l$ and $\mathbf{F}_g$ and multiply them by the temporal feature weight information $\mathcal{W}_{\text{ssm}}$, combining temporal dynamics with spatial structure. A residual connection is then used to combine the merged features with the original input features $\mathbf{F}_f$, producing the final output feature $\mathbf{F}_o$:

$$\mathbf{F}_o = \mathbf{F}_f + \gamma \cdot \text{LN}\big(\mathcal{W}_{\text{ssm}} \odot (\mathbf{F}_l + \mathbf{F}_g)\big), \tag{4}$$

where γ is the scaling factor, LN denotes linear layer, and $\odot$ denotes element-wise multiplication.

In summary, MTS-Mamba's unique dual-path temporal-spatial feature extraction structure effectively captures both multi-scale spatial features and bidirectional temporal dependencies. By introducing forward-backward scanning mechanisms, MTS-Mamba achieves a balance between accuracy in temporal-spatial feature modeling and computational efficiency, significantly enhancing real-time performance in ADAS applications.

C. Multi-task Gated Multimodal Integrator (MGMI)

We propose the Multi-task Gated Multimodal Integrator (MGMI), which introduces task-specific gating mechanisms to dynamically adjust the fusion weights of different modality features, enabling task-driven feature selection and alleviating task conflicts. Specifically, we first concatenate the features $\mathbf{H}_1$, $\mathbf{H}_2 \in \mathbb{R}^{C\times H\times W}$ extracted by MTS-Mamba with the features $\mathbf{H}_3$ from the 3D CNN module along the channel dimension to obtain the initial fused feature. This fused feature undergoes three separate convolution operations to produce queries, keys, and values $\mathbf{Q}, \mathbf{K}, \mathbf{V}$:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathcal{W}_{q,k,v}\big(\text{Concat}(\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3)\big) \in \mathbb{R}^{d\times HW}, \tag{5}$$

where d is the number of channels after projection. The attention matrix is computed by performing a dot product between $\mathbf{Q}$ and $\mathbf{K}$, and multiplying the resulting attention matrix by $\mathbf{V}$ to obtain the weighted attention scores, which are reshaped back to $\mathbb{R}^{d\times H\times W}$.

To adapt to each task's specific focus, we design a multi-gating mechanism. The task-shared features are input into four task-specific gating units, one for each task, which perform convolution, BatchNorm, and Sigmoid operations to calculate attention weights for the multimodal features $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$. A weighted sum of these features is calculated to produce the task-specific feature $\mathbf{F}_r$, where $r \in \{1, 2, 3, 4\}$ represents each task. The task-specific feature $\mathbf{F}_r$ is

$$\mathbf{F}_r = \sum_{i=1}^{3} \mathbf{H}_i \odot \sigma_r^i \left(\text{BN}\big(\text{Conv2D}\big(\text{softmax}(\frac{\mathbf{Q}\cdot\mathbf{K}^{\text{T}}}{\sqrt{d}})\cdot\mathbf{V}\big)\big)\right), \tag{6}$$

where σ_r^i is the i-th gating Sigmoid function for task r, and BN denotes the batch normalization operation. This design optimizes the feature fusion process for each task by dynamically adjusting the importance of each modality's features, allowing the model to extract task-specific features while retaining shared characteristics. By alleviating negative transfer across tasks, MGMI enhances the performance of each task within the multi-task learning framework. Once multimodal features are fused for each task, the multi-task recognition layer applies independent pooling for each task's feature $\mathbf{F}_r$ and inputs them into their respective classifiers, producing the prediction $\hat{y}_r$ for each task.

IV. Experiments

A. Dataset

The AIDE dataset is an open-source collection designed to advance ADAS research, consisting of 2,898 samples of time-series multimodal data, including multi-view images and driver joint positions. The multi-view images are captured from four perspectives: front-view, left-view, right-view, and inside-view. Each sample is annotated with labels for four tasks: driver emotion recognition (DER), driver behavior recognition (DBR), traffic context recognition (TCR), and vehicle behavior recognition (VBR). The dataset is split into training, testing, and validation sets with proportions of 65%, 15%, and 20%, respectively, to ensure robust evaluation.

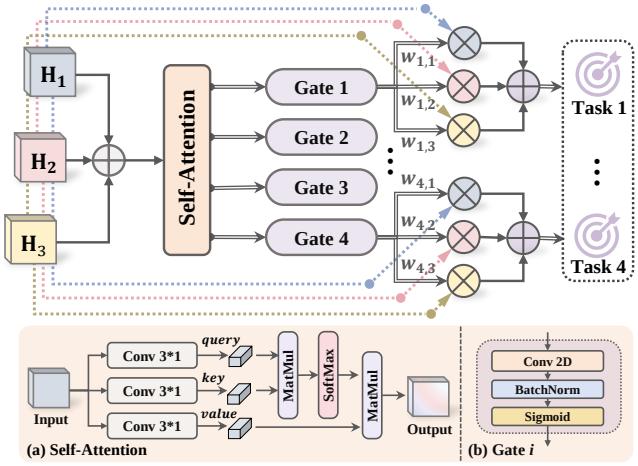

Fig. 4. Structural of the multi-task learning-based gated multimodal feature integrator (MGMI).

B. Data Preprocessing

The preprocessing pipeline follows three sequential steps: extraction of driver facial and body images, data synchronization, and data augmentation. First, driver facial and body regions are cropped using bounding box coordinates from inside-view images. The multimodal data (images and joint positions) is synchronized at 16 frames per second, ensuring temporal alignment of features across modalities. Each input sequence is formed from 16 consecutive frames to capture temporal dependencies. To further augment the dataset, multi-view images are subjected to random horizontal and vertical flips (50% probability), enhancing the diversity of the training set and improving model robustness.

C. Evaluation Metrics

To evaluate the model performance, we use the following key metrics: accuracy (α_{acc}), mean accuracy (β_{macc}), frames per second (FPS), and total parameter count (Params). mAcc represents the average accuracy across multiple tasks, providing an overall performance measure for multi-task learning. It is computed as:

$$\beta_{\text{macc}} = \frac{1}{m}\sum_{r=1}^{m}\alpha_{\text{acc}}^{r}, \tag{7}$$

where and α_{acc}^{r} represents the accuracy of the model on the r-th task.

D. Experimental Details

All experiments were conducted on an NVIDIA L40S GPU with 48GB VRAM. Training was performed for 125 epochs, with stabilization of the outcomes guiding early stopping. The initial learning rate was set to 0.001 and dynamically adjusted: reduced to 0.0005 between epochs 25 and 50, and further decreased to 0.00005 after epoch 50. The batch size was 24, with Stochastic Gradient Descent (SGD) used as the optimization algorithm, featuring a momentum of 0.9 and weight decay of 0.0001 to ensure convergence while mitigating overfitting.

E. Comparison with the State-of-the-art Models

We compared our model with state-of-the-art methods, following the experimental setup from [4], which involves independent feature extraction from multimodal data and multi-view images. Our model consistently outperformed these baseline models across all evaluation metrics, including task-specific α_{acc}, β_{macc}, FPS, and parameter count. With fewer than 6M parameters, our model improved β_{macc} by 3.48%-9.32% and achieved an inference speed of 142.32 FPS, significantly surpassing the real-time requirements of ADAS systems.

These performance improvements are attributed to the model's design optimizations. Specifically, the joint feature extraction across similar modalities reduces model complexity and enhances inter-modal interactions, while the adaptive weighting strategy addresses task-specific feature importance, mitigating negative transfer and improving scalability. These innovations enable our model to achieve superior results in terms of both performance and efficiency, demonstrating its practical applicability for ADAS.

F. Ablation Experiment

We conducted a series of ablation experiments to evaluate the individual contributions of the MTS-Mamba and MGMI modules, as well as their key components, and to investigate the interactive effects between tasks and multimodal data.

1) Ablation experiments on MTS-Mamba and MGMI: We designed three experimental configurations to assess the contributions of the MTS-Mamba and MGMI modules. In the first group, we replaced MTS-Mamba with a simple VGG16 network and MGMI with basic concatenation fusion. The results, shown in Table II, demonstrate significant improvements in β_{macc} by 5.2%-12.04% when both MTS-Mamba and MGMI were used. Specifically, replacing VGG16 with MTS-Mamba improved β_{macc} by 7.66% and increased FPS by 52.89, while reducing parameters by over four times. Similarly, replacing concatenation fusion with MGMI resulted in a 5.2% increase in β_{macc} with only an additional 0.26M parameters. These findings highlight the effectiveness of both MTS-Mamba and MGMI in efficiently extracting temporal-spatial features and mitigating task conflicts in multi-task learning (MTL).

We also evaluated the contribution of the forward-backward temporal scanning mechanisms and the global-local spatial feature extraction module in MTS-Mamba. When either of these components was removed, β_{macc} dropped by 3.47%-3.82%, as shown in Table III. This decline underscores the importance of the synergy between these components in extracting rich spatiotemporal features and synchronizing bidirectional temporal information, providing robust representations for multi-task recognition. Additionally, despite these performance gains, the model maintained its inference speed and did not introduce additional parameters, confirming the efficiency of the MTS-Mamba design.

2) Ablation on Self-Attention and Multi-Gating Mechanisms in MGMI: Finally, we analyzed the role of the self-attention and multi-gating mechanisms in MGMI. Removing

TABLE I

COMPARISON RESULTS WITH OTHER STATE-OF-THE-ART ALGORITHMS ON THE AIDE DATASET. MULTI-VIEW SCENE IMAGES REPRESENT THE FRONT-VIEW, LEFT-VIEW, RIGHT-VIEW, AND INSIDE-VIEW, DRIVER IMAGES REPRESENT THE FACIAL AND BODY IMAGES OF THE DRIVER, AND JOINTS REPRESENT POSTURE AND GESTURE. THE BEST RESULTS ARE MARKED IN **BOLD**, THE SECOND-BEST RESULTS ARE ., AND OUR METHOD IS HIGHLIGHTED WITH . ADDITIONALLY, P(M) REPRESENTS TO THE NUMBER OF PARAMETERS, TE REPRESENTS TEMPORAL EMBEDDING, TRANSE REPRESENTS TRANSFORMER ENCODER [31].

Pattern	Backbone			α_{acc}(%) ↑				β_{macc} (%) ↑	FPS ↑	P(M) ↓
	Multi-view Scene Images	Driver Images	Joints	DER	DBR	TCR	VBR			
2D	VGG16 [32]	VGG16 [32]	3DCNN	69.12	64.57	84.77	74.08	73.15	52.70	127.48
	Res18 [33]	Res18 [33]	3DCNN	68.78	64.33	89.76	78.59	75.37	57.30	107.77
	CMT [34]	CMT [34]	3DCNN	68.75	68.75	93.75	81.38	78.16	61.72	72.33
	GLMDriveNet [12]	GLMDriveNet [12]	3DCNN	71.38	66.57	90.23	77.19	76.34	85.17	78.17
2D + Timing	PP-Res18+TransE [31]	Res18/34 [33]+TransE [31]	MLP+TE	70.83	67.32	90.54	79.97	77.17	-	-
	Res34 [33]+TransE [31]	Res18/34 [33]+TransE [31]	MLP+TE	72.65	67.08	86.63	78.46	76.21	-	-
	Res50 [33]+TransE [31]	Res34/50 [33]+TransE [31]	MLP+TE	70.24	65.65	82.57	77.29	73.94	-	-
	VGG16 [32]+TransE [31]	VGG13/16 [32]+TransE [31]	MLP+TE	71.12	67.15	85.13	78.58	75.50	-	-
	VGG19 [32]+TransE [31]	VGG16/19 [32]+TransE [31]	MLP+TE	69.46	65.48	85.74	77.91	74.65	-	-
3D	3D-Res34 [35]	3D-Res34 [35]	3DCNN	69.13	63.05	87.82	79.31	74.83	12.67	303.10
	MobileNet-V1-3D [36]	MobileNet-V1-3D [36]	ST-GCN	72.23	64.20	88.34	77.83	75.65	33.15	54.05
	MobileNet-V2-3D [37]	MobileNet-V2-3D [37]	ST-GCN	68.47	61.74	86.54	78.66	73.85	12.16	83.78
	ShuffleNet-V1-3D [38]	ShuffleNet-V1-3D [38]	ST-GCN	72.41	68.97	90.64	80.79	78.20	51.29	31.49
	ShuffleNet-V2-3D [39]	ShuffleNet-V2-3D [39]	ST-GCN	70.94	64.04	89.33	78.98	75.82	50.53	35.09
	C3D [40]	C3D [40]	ST-GCN	63.05	63.95	85.41	77.01	72.36	25.62	158.46
	I3D [41]	I3D [41]	ST-GCN	70.94	66.17	87.68	79.81	76.15	-	-
	SlowFast [42]	SlowFast [42]	ST-GCN	72.38	61.58	86.86	78.33	74.79	-	-
	TimeSFormer [43]	TimeSFormer [43]	ST-GCN	74.87	65.18	92.12	78.81	77.75	25.63	158.46
	Video Swin Transformer [44]	Video Swin Transformer [44]	3DCNN	73.44	65.63	93.75	75.00	76.96	11.45	119.80
Ours	**MTS-Mamba**	**MTS-Mamba**	**3DCNN**	**75.00**	**69.31**	**96.29**	**86.11**	**81.68**	**142.32**	**5.99**

TABLE II

ABLATION EXPERIMENT RESULTS OF MTS-MAMBA AND MGMI. "W/" INDICATES THAT THE CORRESPONDING COMPONENT IS USED, WHILE "W/O" DENOTES THAT THE COMPONENT IS NOT USED.

MTS-Mamba	MGMI	α_{acc}(%) ↑				β_{macc} (%) ↑	FPS ↑	P(M) ↓
		DER	DER	TCR	VBR			
w/o	w/o	67.38	58.75	83.06	69.38	69.64	101.83	27.68
w/	w/o	72.13	64.57	92.05	77.16	76.48	**158.84**	**5.73**
w/o	w/	70.54	62.31	90.37	73.84	74.02	89.43	28.12
w/	w/	**75.00**	**69.31**	**96.29**	**86.11**	**81.68**	142.32	5.99

TABLE III

ABLATION EXPERIMENT RESULTS OF THE FORWARD-BACKWARD TEMPORAL SCANNING MECHANISMS AND THE GLOBAL-LOCAL SPATIAL FEATURE EXTRACTION MODULE IN MTS-MAMBA.

Global-Local	Dual-path Scanning	α_{acc}(%) ↑				β_{macc} (%) ↑	FPS ↑	P(M) ↓
		DER	DER	TCR	VBR			
w/	w/o	73.15	66.85	91.21	80.22	77.86	146.59	**5.99**
w/o	w/	72.78	67.30	92.11	80.65	78.21	**148.84**	**5.99**
w/	w/	**75.00**	**69.31**	**96.29**	**86.11**	**81.68**	142.32	**5.99**

TABLE IV

ABLATION EXPERIMENT RESULTS OF THE SELF-ATTENTION AND MULTI-GATING MECHANISMS IN MGMI.

Self-Attention	Multi-gating Mechanism	α_{acc}(%) ↑				β_{macc} (%) ↑
		DER	DER	TCR	VBR	
w/o	w/o	72.13	64.57	92.05	77.16	76.48
w/	w/o	74.60	65.84	93.29	84.75	79.62
w/o	w/	73.53	64.92	91.99	82.39	78.21
w/	w/	**75.00**	**69.31**	**96.29**	**86.11**	**81.68**

either of these components led to a noticeable decline in β_{macc}, especially when the multi-gating mechanism was removed (Table IV). This result is expected, as the multi-gating mechanism dynamically adjusts the fusion weights of each modality's features according to the task's requirements, enabling selective emphasis on the most relevant modality for each task. This adjustment alleviates feature conflicts in MTL and significantly improves the model's performance across multiple tasks. The experimental results reinforce the importance of both the self-attention and multi-gating mechanisms in enhancing feature fusion for multi-task learning.

3) Ablation experiments between different tasks: To explore the advantages of MTL and the interactions between different tasks, we designed a series of ablation experiments. The four tasks were grouped into two dimensions: driver state recognition (DER and DBR) and traffic environment recognition (TCR and VBR). We performed two sets of experiments (Table V): we retained only the driver state recognition tasks (DER and DBR), excluding the traffic environment tasks (TCR and VBR). The results showed a drop in α_{acc} for DER and DBR by 1.86%-2.13%. In the second set, we retained only the traffic environment recognition tasks (TCR and VBR), excluding the driver state tasks (DER and DBR), which resulted in a more significant

TABLE V

ABLATION EXPERIMENTAL RESULTS FOR DRIVER STATE RECOGNITION TASKS (I.E., DER, DBR) AND TRAFFIC ENVIRONMENT RECOGNITION TASKS (I.E., TCR, VBR).

Task		$\alpha_{\text{acc}}(\%) \uparrow$			
Driver States	Traffic Environment	DER	DER	TCR	VBR
w/	w/o	73.14	67.18	-	-
w/o	w/	-	-	91.47	80.25
w/	w/	**75.00**	**69.31**	**96.29**	**86.11**

TABLE VI

RESULTS OF THE ABLATION EXPERIMENTS ON MULTIMODAL DATA.

Vehicle-exterior Images	Vehicle-interior Images	Joints	$\alpha_{\text{acc}}(\%) \uparrow$				β_{macc} $(\%) \uparrow$
			DER	DER	TCR	VBR	
✔			69.78	60.27	92.71	80.68	75.86
	✔		71.23	61.49	84.86	72.03	72.40
		✔	70.39	65.53	73.33	60.74	67.50
✔	✔	✔	**75.00**	**69.31**	**96.29**	**86.11**	**81.68**

α_{acc} drop of 4.82%-5.86%. These results demonstrate that the tasks within the MTL framework benefit from significant synergies. Jointly learning both driver state and traffic environment recognition tasks enhances the model's overall accuracy and generalization capability.

4) Ablation experiments on multimodal data: This subsection validates the independent contributions of each modality through ablation experiments. We categorize the modalities into three groups: vehicle-exterior images (front-view, left-view, right-view), vehicle-interior images (inside-view, driver's facial and body images), and joint data (posture and gesture). We trained the model with each data group separately, and the results are shown in Table VI.

The results demonstrate that models trained with a single modality perform worse than the multimodal model, with a drop of 5.82% to 14.18% in β_{macc}. This confirms the crucial role of multimodal data in ADAS-related tasks. Further analysis reveals that different modalities benefit different tasks: vehicle-exterior images improve accuracy for TCR and VBR, while vehicle-interior images and joint data enhance DER and DBR accuracy. This variation arises because each modality expresses different information—vehicle-exterior images reflect road conditions, while vehicle-interior images and joint data capture the driver's behavior and facial expressions. These findings validate the necessity of our MGMI design, which adaptively adjusts the weights of different modalities to alleviate negative transfer and improve task performance in MTL scenarios.

V. CONCLUSION

This paper introduces a TEM^3-Learning (Time-Efficient Multimodal Multi-Task Learning) framework designed to recognize driver emotion, behavior, traffic context, and vehicle behavior simultaneously. At its core, TEM^3-Learning integrates two key components: MTS-Mamba, which efficiently captures temporal-spatial features from multi-view sequential images, and MGMI, which adaptively adjusts the weights of modality features for each task using a multi-gate mechanism. This design alleviates negative transfer between tasks, optimizing performance across multiple recognition tasks. Experimental results on the AIDE dataset demonstrate that TEM^3-Learning achieves superior performance in all four recognition tasks, with an inference speed exceeding the baseline models, while maintaining fewer than 6 million parameters. These findings highlight the efficiency, scalability, and practical applicability of TEM^3-Learning in real-time ADAS systems. We believe that TEM^3-Learning and its core components offer a valuable contribution to multimodal multi-task learning in ADAS, paving the way for the development of more efficient and robust algorithms in this field.

REFERENCES

[1] Y. Gong, J. Lu, W. Liu, Z. Li, X. Jiang, X. Gao, and X. Wu, "Sifdrivenet: Speed and image fusion for driving behavior classification network," *IEEE Transactions on Computational Social Systems*, 2023.

[2] X. Zhang, Y. Gong, J. Lu, Z. Li, S. Li, S. Wang, W. Liu, L. Wang, and J. Li, "Oblique convolution: A novel convolution idea for redefining lane detection," *IEEE Transactions on Intelligent Vehicles*, 2023.

[3] Y. Qian, J. M. Dolan, and M. Yang, "Dlt-net: Joint detection of drivable areas, lane lines, and traffic objects," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4670–4679, 2019.

[4] D. Yang, S. Huang, Z. Xu, Z. Li, S. Wang, M. Li, Y. Wang, Y. Liu, K. Yang, Z. Chen *et al.*, "Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 459–20 470.

[5] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2801–2810.

[6] S. Chowdhuri, T. Pankaj, and K. Zipser, "Multinet: Multi-modal multi-task learning for autonomous driving," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1496–1504.

[7] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "Yolop: You only look once for panoptic driving perception," *Machine Intelligence Research*, vol. 19, no. 6, pp. 550–562, 2022.

[8] Y. Xing, C. Lv, D. Cao, and E. Velenis, "Multi-scale driver behavior modeling based on deep spatial-temporal representation for intelligent vehicles," *Transportation research part C: emerging technologies*, vol. 130, p. 103288, 2021.

[9] S. Liu, Y. Liang, and A. Gitter, "Loss-balanced task weighting to reduce negative transfer in multi-task learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9977–9978.

[10] W. Choi, M. Shin, H. Lee, J. Cho, J. Park, and S. Im, "Multi-task learning for real-time autonomous driving leveraging task-adaptive attention generator," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 732–14 739.

[11] C. Guo, H. Liu, J. Chen, and H. Ma, "Temporal information fusion network for driving behavior prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9415–9424, 2023.

[12] W. Liu, Y. Gong, G. Zhang, J. Lu, Y. Zhou, and J. Liao, "Glmdrivenet: Global–local multimodal fusion driving behavior classification network," *Engineering Applications of Artificial Intelligence*, vol. 129, p. 107575, 2024.

[13] L. Mou, Y. Zhao, C. Zhou, B. Nakisa, M. N. Rastgoo, L. Ma, T. Huang, B. Yin, R. Jain, and W. Gao, "Driver emotion recognition with a hybrid attentional multimodal fusion framework," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2970–2981, 2023.

[14] M. Hao, Z. Zhang, L. Li, K. Dong, L. Cheng, P. Tiwari, and X. Ning, "Coarse to fine-based image–point cloud fusion network for 3d object detection," *Information Fusion*, vol. 112, p. 102551, 2024.

[15] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 850–17 859.

[16] W. Liu, J. Lu, J. Liao, Y. Qiao, G. Zhang, J. Zhu, B. Xu, and Z. Li, "Fmdnet: Feature-attention-embedding-based multimodal-fusion driving-behavior-classification network," *IEEE Transactions on Computational Social Systems*, 2024.

[17] Y. Huang, W. Liu, Y. Li, L. Yang, H. Jiang, Z. Li, and J. Li, "Mfe-ssnet: Multi-modal fusion-based end-to-end steering angle and vehicle speed prediction network," *Automotive Innovation*, pp. 1–14, 2024.

[18] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1930–1939.

[19] Y. Gong, J. Lu, J. Wu, and W. Liu, "Multi-modal fusion technology based on vehicle information: A survey," *arXiv preprint arXiv:2211.06080*, 2022.

[20] Y. Gan, W. Liu, J. Gan, and G. Zhang, "A segmentation method based on boundary fracture correction for froth scale measurement," *Applied Intelligence*, pp. 1–22, 2024.

[21] X. Shi, Z. Yin, G. Han, W. Liu, L. Qin, Y. Bi, and S. Li, "Bssnet: A real-time semantic segmentation network for road scenes inspired from autoencoder," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[22] X. Wang, K. Huang, X. Zhang, H. Sun, W. Liu, H. Liu, J. Li, and P. Lu, "Path planning for air-ground robot considering modal switching point optimization," in *2023 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2023, pp. 87–94.

[23] Z. Li, T. Zhang, M. Zhou, D. Tang, P. Zhang, W. Liu, Q. Yang, T. Shen, K. Wang, and H. Liu, "Mipd: A multi-sensory interactive perception dataset for embodied intelligent driving," *arXiv preprint arXiv:2411.05881*, 2024.

[24] D. Zhou, H. Liu, H. Ma, X. Wang, X. Zhang, and Y. Dong, "Driving behavior prediction considering cognitive prior and driving context," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 2669–2678, 2020.

[25] Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, and X. Bai, "Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 7, pp. 8324–8341, 2022.

[26] W.-H. Li and H. Bilen, "Knowledge distillation for multi-task learning," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 163–176.

[27] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 675–684.

[28] J. Cui, J. Du, W. Liu, and Z. Lian, "Textnerf: A novel scene-text image synthesis method based on neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 272–22 281.

[29] M. Gao, J.-Y. Li, C.-H. Chen, Y. Li, J. Zhang, and Z.-H. Zhan, "Enhanced multi-task learning and knowledge graph-based recommender system," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10 281–10 294, 2023.

[30] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[34] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 12 175–12 185.

[35] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.

[36] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[38] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856.

[39] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 116–131.

[40] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[41] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[42] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6202–6211.

[43] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *International Conference on Machine Learnin (ICML)*, 2021, p. 4.

[44] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.