

Spatial-Temporal Perception with Causal Inference for Naturalistic Driving Action Recognition

Qing Chang^{1*}, Wei Dai^{2*}, Zhihao Shuai^{3*}, Limin Yu², Yutao Yue^{3†}

¹School of Mechanical Engineering, Nanjing University of Science and Technology

²School of Advanced Technology, Xi'an Jiaotong-Liverpool University

³The Hong Kong University of Science and Technology (Guangzhou)

Abstract—Naturalistic driving action recognition is essential for vehicle cabin monitoring systems. However, the complexity of real-world backgrounds presents significant challenges, and previous approaches have struggled with practical implementation due to their limited ability to observe subtle behavioral differences and effectively learn inter-frame features from video. In this paper, we propose a novel Spatial-Temporal Perception (STP) architecture that emphasizes both temporal information and spatial relationships between key objects, incorporating a causal decoder to perform behavior recognition and temporal action localization. Without requiring multimodal input, STP directly extracts temporal and spatial distance features from RGB video clips. Subsequently, these dual features are jointly encoded by maximizing the expected likelihood across all possible permutations of the factorization order. By integrating temporal and spatial features at different scales, STP can perceive subtle behavioral changes in challenging scenarios. Additionally, we introduce a causal-aware module to explore relationships between video frame features, significantly enhancing detection efficiency and performance. We validate the effectiveness of our approach using two publicly available driver distraction detection benchmarks. The results demonstrate that our framework achieves state-of-the-art performance.

Index Terms—driver action recognition, causal inference, car cabin monitoring

I. INTRODUCTION

Naturalistic driving action recognition (DAR) is a critical component of vehicle cabin monitoring systems. Its primary objectives are distracted behavior detection and temporal action localization (TAL) within untrimmed video sequences, both of which are vital for improving driving safety and fostering effective driver-vehicle interaction.

Recent advancements in DAR have been propelled by the powerful representation capabilities of deep learning [1]. Several approaches build on general human action recognition backbones, leveraging 3D convolutional neural networks (CNNs) [4] and vision transformers [24]. Among these, the temporal shift module (TSM) [13] has demonstrated effectiveness in learning features from adjacent frames.

Despite this progress, DAR remains a challenging task due to the complex nature of the vehicle cabin environment, which provides limited distinguishing features. As shown in Fig. 1, drivers often exhibit highly similar movements of body parts

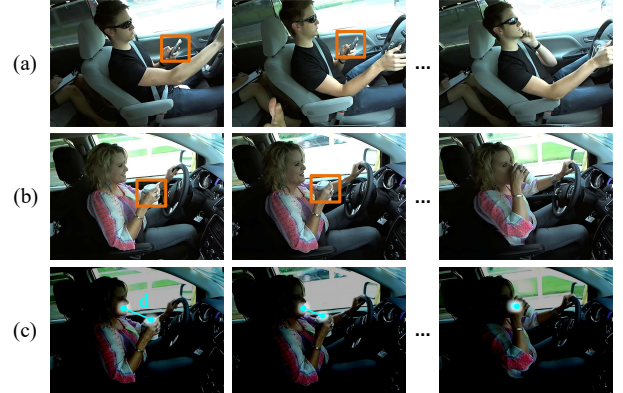


Fig. 1. Illustrations of challenging cases in driving action recognition. (a) Calling and (b) drinking scenes, where the objects are partially visible and the lighting conditions are unstable. (c) Variations in the distance d between key points assist in identifying behavior categories and temporal localization.

(e.g., eating and drinking), which can easily confuse detectors. Additionally, variable lighting conditions in the cabin and the duration of input video clips pose further challenges in modeling long-sequence feature relationships.

Several studies have sought to address these challenges. Khan et al. [7] utilized depth and infrared inputs using a late fusion method to improve the robustness of driving behavior detection. Ma et al. [16] proposed a multi-scale attention module for multi-view image fusion, while Jiang et al. [9] developed a multi-camera DAR model that trains a single-camera feature extractor to boost performance. However, these methods are limited by the requirement for additional input types and depend on single-mode classification pipelines, which reduce their practicality in real-world hardware environments and compromise efficiency. Additionally, they often neglect the temporal correlations between frames.

After thorough analysis, we argue that two critical features from video merit attention: temporal information and the spatial distance between interest objects. Temporal information is directly derived from video clips, providing fine-grained visual features essential for action recognition. Furthermore, as illustrated in Fig. 1(c), changes in the distance between a driver's mouth and a cup offer cues for identifying the start and end of actions. By aligning and fusing these two feature types, the model can better focus on key regions to perceive

*These authors contributed equally.

†Corresponding author: yutaoyue@hkust-gz.edu.cn

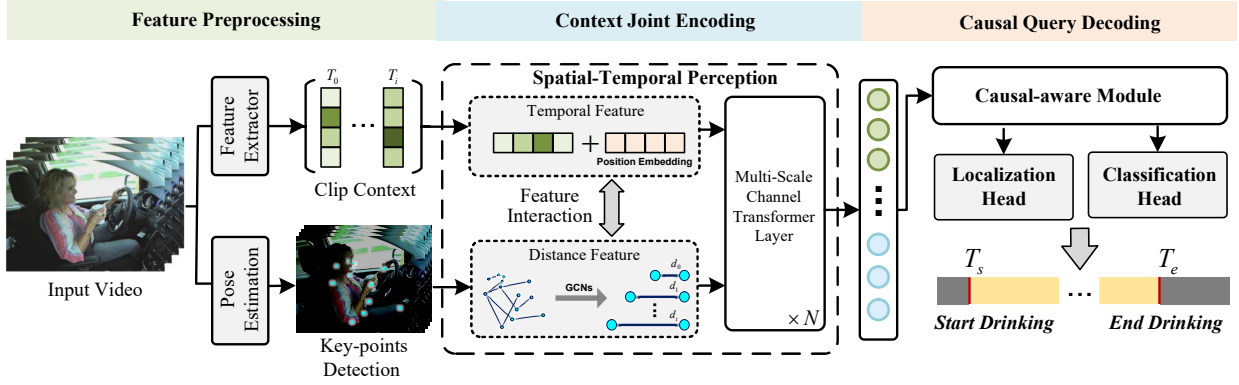


Fig. 2. The overall architecture of STP.

the action, thereby mitigating irrelevant interference.

To this end, we present a novel Spatial-Temporal Perception (STP) network that emphasizes both temporal information and spatial relationships between interest objects, incorporating a causal decoder to perform behavior recognition and temporal action localization. Without relying on multi-view and multimodal input, STP directly extracts temporal and spatial distance features from RGB video clips. These dual features are jointly encoded by maximizing the expected likelihood across all possible permutations of the factorization order. By integrating temporal and spatial features, STP is capable of detecting subtle behavioral changes in challenging scenarios. Furthermore, we introduce a causal-aware module to analyze relationships between video frame features, significantly improving detection efficiency and performance. We validated the effectiveness of our approach using two publicly available driver distraction detection benchmarks: Drive&Act and SynDD2. The results demonstrate that our framework achieves state-of-the-art performance in both driver action recognition and temporal action localization tasks.

II. METHODOLOGY

A. Overview

The overall architecture of STP is illustrated in Fig. 2. Given a video with T frames, denoted as $X \in \mathbb{R}^{T \times 3 \times H \times W}$, the STP aims to integrate the temporal features of video clips with the spatial relationships between key points to enhance driver action recognition and temporal localization. The video clip is first processed in parallel by two heads to extract the clip context and key point positions. These two outputs are then combined, where the clip context is aggregated with position embeddings to generate temporal features, and key points are refined into distance features for each frame using graph convolutional networks (GCNs) [8]. These two types of features are interactively fused to align in space and passed to a stacked multi-scale channel transformer for context encoding. The proposed causal-aware module further explores the relationships between feature sequences. Finally, these hybrid features are decoded by the localization and classification heads to accurately identify the behavior pattern and the start and end frames of action.

B. Context Joint Encoding

In the feature extraction stage, two lightweight extraction networks are employed to obtain dual features. However, the dense temporal features are not aligned with the sparse distance features between nodes extracted by GCNs. To address this misalignment, we first introduce an interaction mechanism to spatially align the dual feature sets. These aligned features are then fed into stacked multi-scale channel transformer layers for fusion and calibration, enabling the integration of both global and local information.

Dual Feature Interaction. Feature interaction transfers and aligns features between different modalities. Let the T -frame dual feature clips be represented as $X^p = \{x_t^p\}_{t=1}^T$ and $X^d = \{x_t^d\}_{t=1}^T$, where $x_t^p, x_t^d \in \mathbb{R}^{C \times L}$ denote the temporal and distance features at timestamp t , respectively. The dual features are updated by shifting the last k feature channels of each modality as follows:

$$\hat{x}_t^p = \text{MLP} \left(x_t^p[: -k], x_t^d[-k :] \right), \quad (1)$$

$$\hat{x}_t^d = \text{MLP} \left(x_t^d[: -k], x_t^p[-k :] \right), \quad (2)$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation, and MLP refers to a fully connected layer. This process incurs minimal computational cost. As a result, dual-feature interaction efficiently aligns and integrates information across modalities. **Multi-Scale Encoder.** We adopt a stacked multi-scale channel transformer layer within our recurrence mechanism to facilitate the reuse of hidden states from preceding segments. For a longer sequence F obtained via dual-feature interaction, consider extracting two segments $\tilde{S} = F_{1:t}$ and $S = F_{t+1:2t}$ for illustration. We process the initial segment \tilde{S} and retain the resultant content representations $\tilde{H}^{(m)}$ for each layer m . Let $Q = H^{(m-1)}$ and $K, V = [\tilde{H}^{(m-1)}, H^{(m-1)}]$. When processing the subsequent segment, the attention update, integrating memory, can be formulated as follows:

$$H^{(m)} = \text{Softmax} \left(\frac{QK^T}{\sqrt{D}} \right) V, \quad (3)$$

where D denotes the embedding dimension. Consequently, once the representations $\tilde{H}^{(m-1)}$ have been obtained, the attention update operates independently of the variable \tilde{S} .

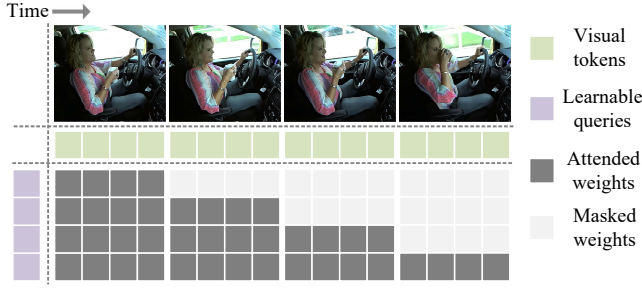


Fig. 3. Causal-aware module incrementally exposes video frames to learnable queries to decouple spatial and temporal features.

C. Causal Query Decoding

Causal-aware Module. To ensure the query attends equally to the embeddings of each time frame and fully explores the causal relationships between video frames, we propose a causal-aware module based on cross-attention masks. Specifically, the output of the module for video embeddings is computed as follows:

$$y_i = \frac{\sum_j M_{ij} \exp(Q_i K^T(x_j)) V(x_j)}{\sum_j M_{ij} \exp(Q_i K^T(x_j)) \mathbf{1}_L}, \quad (4)$$

where M_{ij} denotes the mask for the i -th query Q_i of the j -th frame x_j . As current time step t is typically smaller than the number of tokens n , $M_{ij} = 1$ if $i \geq j \lfloor \frac{n}{t} \rfloor$, otherwise $M_{ij} = 0$. The $\mathbf{1}_L = [1, \dots, 1]^T \in \mathbb{R}^{L \times 1}$ is a vector of ones. We illustrate the masking process of our module in Fig. 3. This approach ensures that initial queries focus on early visual embeddings, while final queries can access embeddings from various time frames to capture causal relationships across time.

Prediction Head. The prediction head consists of a classification head and a localization head. A feed-forward network (FFN) [23] is used to predict parameters, and the prediction head is formulated as:

$$[T_s, T_e] = \text{FFN}_{reg}(\mathbf{Q}), \quad (5)$$

$$\mathcal{C} = \text{FFN}_{cls}(\mathbf{Q}), \quad (6)$$

where T_s, T_e represent the start and end frames of the action, and \mathcal{C} denotes the predicted action category.

Total Loss. Given matched ground truth labels for the prediction queries, we calculate the corresponding loss for each matched pair. The overall loss of our model includes both classification loss and regression loss:

$$\mathcal{L}_{total} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_{reg} \mathcal{L}_{reg}, \quad (7)$$

where \mathcal{L}_{cls} is the focal loss with $\gamma = 2.0$ and $\alpha = 0.25$. \mathcal{L}_{reg} is the smooth- l_1 loss for TAL.

III. EXPERIMENTS AND RESULTS

A. Datasets and Evaluation Metrics

Drive&Act [18] is widely used for driver activity recognition tasks. It contains 9.6 million frames across three modalities (RGB, IR, and depth) and five different camera views. The dataset provides three levels of activity labels: action units,

TABLE I
COMPARISON WITH POPULAR METHODS ON DRIVE&ACT.

Method	Modality	Mean-1(%) \uparrow	Top-1(%) \uparrow
ResNet [6]	IR, Depth	51.08	56.43
UniFormerV2 [10]	RGB, IR, Depth	61.58	78.63
MDBU (I3D) [21]	IR, NIR	62.02	76.91
DFS [12]	IR, Depth	63.12	77.61
TSM [13]	IR	59.81	67.75
TransDARC [20]	RGB	60.10	76.17
UniFormerV2 [10]	RGB	61.79	76.71
STP (Ours)	RGB	64.82	79.32

TABLE II
COMPARISON WITH POPULAR METHODS ON SYND2.

Method	Multi-View	Setting	AO-Score \uparrow
M2DAR [17]	\checkmark	Right, Dashboard	0.5921
MCPRL [26]	\checkmark	Right, Rear, Dashboard	0.6080
SKKU [19]	\checkmark	Right, Rear, Dashboard	0.7798
APC [11]	\times	Dashboard	0.7046
AMA [25]	\times	Right	0.7459
STP (Ours)	\times	Right	0.7923

fine-grained activities, and coarse tasks. In this paper, we focus on the fine-grained RGB modality from the top-right view.

SynDD2 [22] includes IR and RGB videos, along with annotation files, collected from three in-vehicle cameras located at the dashboard, rearview mirror, and top-right corner of the window. The dataset covers two types of activities: distracted activities and gaze zones, each with and without appearance obstructions such as hats or sunglasses.

Evaluation Metrics. We follow the official evaluation metrics for Drive&Act, using Mean-1 Accuracy (average per-class accuracy) as the primary metric, and Top-1 Accuracy for implementation assessment. For SynDD2, we evaluate temporal action localization and recognition performance using the average overlap score (AO-Score), which is defined as follows:

$$os(p, g) = \frac{\max(\min(ge, pe) - \max(gs, ps), 0)}{\max(ge, pe) - \min(gs, ps)}, \quad (8)$$

where gs and ge represent the start and end times of the ground-truth activity g , respectively. The variable p denotes the best predicted activity of the same category as g , while os refers to the highest overlap. The overlap between g and p is defined as the ratio of the intersection time to the union time of the two activities. After matching each ground truth activity in order of their start times, any unmatched ground truth activities or unmatched predicted activities will be assigned an overlap score of 0.

B. Implementation Details

We utilize the pre-trained VideoMAEv2 [24] and OpenPose [3] models as the backbones for video feature extraction and spatial pose estimation, respectively. Following [5], the input video is sampled with a temporal stride of 8, each frame is resized to 224×224 , and only 13 key points are used per frame. The Multi-Scale Encoder consists of 6 layers, with 4 heads and 256-dimensional embeddings. In the training stage,

TABLE III
EFFECTS OF EACH COMPONENT IN OUR METHOD.

Spatial Feature	Temporal Feature	Spatial-Temporal Feature	Causal-aware Model	AO-Score \uparrow
✓				0.7223
	✓			0.7298
		✓		0.7643
		✓	✓	0.7923

TABLE IV
THE COMPARISON OF THE MODEL EFFICIENCY RESULTS.

Modality	Methods	Latency(ms) \downarrow	#Param \downarrow
Dual	UniFormerV2 [10]	33.0	47.2M
Dual	DFS [12]	28.0	38.8M
Single	TSM [13]	15.0	25.3M
Single	I3D [21]	18.3	28.0M
Single	STP (Ours)	14.2	23.7M

we use AdamW [15] optimizer with an initial learning rate of $1e-3$ and cosine decay learning rate strategy [14] with power set to 0.9. During inference, the initial predictions are screened by SoftNMS [2] with a threshold of 0.2. The weights λ_{cls} and λ_{reg} are set as 1 and 1.5, respectively. All experiments are performed on GeForce RTX 3090 GPU.

C. Quantitative Results

Table I shows the results of our method on the Drive&Act test dataset. We compare popular single-model and multi-modal driving action recognition methods. STP significantly outperforms all methods, achieving a Mean-1 accuracy of 64.82% and a Top-1 accuracy of 79.32%. This shows that our method can effectively achieve high-precision driver action recognition even without additional information input.


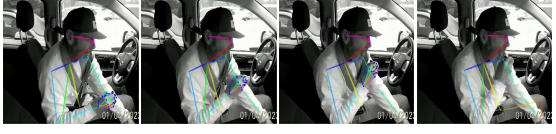

Table II shows the average overlap score on the SynDD2 dataset. In this table, we categorize and compare the methods using different camera angles. Without the need for complex multi-view fusion, our approach achieved a state-of-the-art performance with 0.7923 AO-Score, demonstrating the effectiveness of our proposed Spatial-Temporal Perception. It is worth noting that our method only relies on the RGB input of the camera on the right side of the driver, which greatly reduces the hardware cost of the actual scene.

D. Ablation Study and Visualization

To validate the effectiveness of our STP model, we conducted several ablation experiments on the SynDD2 validation dataset. Furthermore, we compared the efficiency of the current popular method and visualized the prediction results to better emphasize the advantages of our approach.

Ablation Study. We present the results of the ablation studies in Table III. We progressively add the spatial-temporal perception structure and the causal-aware module, and report the corresponding AO-Score. The results indicate that using spatial or temporal features alone provides only limited improvements. In contrast, the proposed spatial-temporal perception

TABLE V
SAMPLE VISUALIZATIONS OF THE PROCESS OF KEYPOINT DETECTION.

Input	Results
	Calling
	Eating
	Drinking

significantly enhances the AO-Score. Furthermore, the causal-aware module effectively captures relationships between video frames, leading to additional performance gains.

Efficiency comparison. Model efficiency is critical for real-time driver monitoring systems. We further evaluate the efficiency of the model in terms of latency and parameter size, as shown in Table IV. For a fair comparison, we categorize the current popular methods into dual and single modality inputs, ensuring consistent input cropping. The results demonstrate that our method not only achieves superior performance but also retains the efficiency benefits of lower latency and a reduced parameter count typical of single-modality inputs.

Results Visualization. As shown in Table V, we further visualize several challenging cases and their corresponding results during the keypoint detection stage. These examples clearly demonstrate that changes in the distance between key points (such as between the fingers and mouth) provide valuable prior knowledge, enabling the model to make accurate inferences. Even when actions appear similar, our method can accurately distinguish between the driver's Calling and Eating actions and predict the precise start and end times of these actions.

IV. CONCLUSION

In this paper, we introduce a novel Spatial-Temporal Perception (STP) architecture designed to enhance action recognition and temporal action localization by capturing both the temporal dynamics and spatial relationships between key objects. Unlike multimodal approaches, STP directly extracts temporal and spatial distance features from RGB video clips, encoding these dual features by optimizing the likelihood across various factorization orders. This integration allows STP to detect subtle behavioral changes, even in complex scenarios. Furthermore, the inclusion of a causal-aware module improves detection efficiency by exploring the relationships between video frame features. Validated on two publicly available driver distraction detection benchmarks, our approach achieves state-of-the-art performance, highlighting its effectiveness and potential for broader applications.

REFERENCES

- [1] V. A. Adewopo, N. Elsayed, Z. ElSayed, M. Ozer, A. Abdelgawad, and M. Bayoumi. A review on action recognition for accident detection in smart city transportation systems. *Journal of Electrical Systems and Information Technology*, 10(1):57, 2023. 1
- [2] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Softnms—improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 4
- [3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3
- [4] H. Chen. Skateboardai: The coolest video action recognition for skateboarding (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16184–16185, 2023. 1
- [5] X. Dong, R. Zhao, H. Sun, D. Wu, J. Wang, X. Zhou, J. Liu, S. Cui, and Z. He. Multi-attention transformer for naturalistic driving action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5435–5441, 2023. 3
- [6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [7] S. S. Khan, Z. Shen, H. Sun, A. Patel, and A. Abedi. Supervised contrastive learning for detecting anomalous driving behaviours from multimodal videos. In *2022 19th Conference on Robots and Vision (CRV)*, pages 16–23. IEEE, 2022. 1
- [8] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2
- [9] J. Kuang, W. Li, F. Li, J. Zhang, and Z. Wu. Mifi: Multi-camera feature integration for robust 3d distracted driver activity recognition. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 1
- [10] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao. Uniformerv2: Unlocking the potential of image vits for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1632–1643, 2023. 3, 4
- [11] R. Li, C. Wu, L. Li, Z. Shen, T. Xu, X.-j. Wu, X. Li, J. Lu, and J. Kittler. Action probability calibration for efficient naturalistic driving action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5277, 2023. 3
- [12] D. Lin, P. H. Y. Lee, Y. Li, R. Wang, K.-H. Yap, B. Li, and Y. S. Ngim. Multi-modality action recognition based on dual feature shift in vehicle cabin monitoring. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6480–6484. IEEE, 2024. 3, 4
- [13] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 1, 3, 4
- [14] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4
- [15] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4
- [16] Y. Ma, V. Sanchez, S. Nikan, D. Upadhyay, B. Atote, and T. Guha. Real-time driver monitoring systems through modality and view analysis. *arXiv preprint arXiv:2210.09441*, 2022. 1
- [17] Y. Ma, L. Yuan, A. Abdelraouf, K. Han, R. Gupta, Z. Li, and Z. Wang. M2dar: Multi-view multi-scale driver action recognition with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5287–5294, 2023. 3
- [18] 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*, pages 2801–2810, 2019. 3
- [19] H.-H. Nguyen, C. D. Tran, L. H. Pham, D. N.-N. Tran, T. H.-P. Tran, D. K. Vu, Q. P.-N. Ho, N. D.-M. Huynh, H.-M. Jeon, H.-J. Jeon, et al. Multi-view spatial-temporal learning for understanding unusual behaviors in untrimmed naturalistic driving videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7144–7152, 2024. 3
- [20] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen. Transdarc: Transformer-based driver activity recognition with latent space feature calibration. in 2022 ieee. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 278–285. 3
- [21] A. Roitberg, K. Peng, Z. Marinov, C. Seibold, D. Schneider, and R. Stiefelhagen. A comparative analysis of decision-level fusion for multimodal driver behaviour understanding. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1438–1444. IEEE, 2022. 3, 4
- [22] M. Shaiqur Rahman, J. Wang, S. Velipasalar Gursoy, D. Anastasiu, S. Wang, and A. Sharma. Synthetic distracted driving (syndd2) dataset for analyzing distracted behaviors and various gaze zones of a driver. *arXiv e-prints*, pages arXiv–2204, 2022. 3
- [23] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3
- [24] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 1, 3
- [25] T. Zhang, Q. Wang, X. Dong, W. Yu, H. Sun, X. Zhou, A. Zhen, S. Cui, D. Wu, and Z. He. Augmented self-mask attention transformer for naturalistic driving action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7108–7114, 2024. 3
- [26] W. Zhou, Y. Qian, Z. Jie, and L. Ma. Multi view action recognition for distracted driver behavior localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5375–5380, 2023. 3