VNU Journal of Science: Com. Science & Com. Eng., Vol. 42, No.xx (2026) xx–xx

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

MAE4DAR: Masked AutoEncoder for Driver Action Recognition

Quang-Tung Nguyen[1], Thuy-Binh Nguyen[2*], Hong-Quan Nguyen[3], Thi-Lan Le[1]

[1] *SigM Lab, School of Electrical and Electronic Engineering (SEEE), HUST, Hanoi, Vietnam.*
[2] *University of Transport and Communications, Hanoi, VietNam.*
[3] *Viet-Hung Industrial University, Hanoi, Vietnam.*

Received 18th March 2026
Revised 02nd May 2026; Accepted 29th June 2026

Abstract: Driver Action Recognition (DAR) plays an important role in intelligent vehicle systems and road safety applications. DAR is a subfield of Human Action Recognition (HAR) that identifies actions in videos. Besides the challenges inherited from HAR, DAR also faces severe occlusion issues in confined environments such as vehicle cabins. In this paper, we introduce a Transformer-based framework for DAR, namely MAE4DAR, that leverages the representation learning capability of VideoMAE V2 as a video encoder. The contribution of this work is twofold. First, we introduce a unified pipeline that handles both isolated and continuous recognition scenarios on the same VideoMAE-based backbone. In isolated recognition setting, some experiments are conducted on two datasets for driver action recognition that are UTCDriverAct and Driver Action Dataset (DAD). On UTCDriverAct, MAE4DAR framework achieves strong performance, reaching perfect recognition accuracy in four of the six driver action classes. In continuous setting, the proposed framework yields a frame-wise accuracy of 92.7% on both views, with overlap scores of 0.72 and 0.69 for front-view and rear-view data in UTCDriverAct dataset, respectively. Second, we propose multi-view fusion at score level to combine prediction scores obtained from independently trained single-view models. The experimental results indicate that multi-view fusion consistently improves recognition performance, achieving an overlap score of 0.75 and a frame-wise accuracy of 93.5%, outperforming both single-view settings. Furthermore, we analyze the impact of multi-view fusion on both recognition performance and computational cost, demonstrating that its performance gains justify the additional inference overhead, which can be mitigated through parallel processing.

Keywords: Driver Action Recognition, Human Action Recognition, VideoMAE encoder, Transformer-based Action Recognition.

*Corresponding author.
E-mail address: thuybinh_ktdt@utc.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.7153

1. Introduction

Traffic accidents remain a critical issue for countries worldwide, particularly in the current context of the rapidly increasing number of vehicles. One of the leading causes of serious and fatal road accidents is driver distraction during driving [1]. This indicates the necessity of developing an automated system which can detect and recognize distracted driver actions, known as driver action recognition (DAR). The primary goal of this system is to provide useful warnings relied on recognition results for both drivers and authorities, thereby reducing traffic accidents and improving road safety. In the literature, most existing studies on DAR are categorized into two primary approaches that are wearable sensor-based and vision-based methods. For the first approach, several wearable sensors are attached to the human body to collect various biomedical signals for identifying the driver's distracted states. However, this approach primarily focuses on detecting the driver's drowsiness or fatigue rather than recognizing a broader range of distracted driving actions. Furthermore, attaching sensors to the human body may cause inconvenience for users, particularly when they are required to wear these sensors for extended periods. Consequently, the majority of recent DAR studies have focused on vision-based approaches, where visual information is captured and processed by automated recognition systems to analyze driver behaviors.

In vision-based approaches, existing studies typically investigate DAR under two main settings that are isolated recognition and continuous recognition. While isolated recognition methods focus on trimmed videos consisting of a single driver action and are generally formulated as a classification task. Inversely, continuous recognition deals with untrimmed videos where multiple driver actions may occur. Accordingly, continuous recognition methods are required not only to determine which action is taking place but also to identify the temporal boundaries of each action, including its start and end times. Although isolated recognition setting does not fully reflect real-world driving scenarios, the results obtained from this setting often serve as a fundamental basis for the development of continuous recognition approach.

In the field of pattern recognition, driver action recognition (DAR) is regarded as a subfield of human action recognition (HAR). Consequently, DAR inherits several challenges commonly encountered in HAR, such as background clutter, similarities between actions of different classes, and variations within the same class. Moreover, DAR has to face additional difficulties specific to the driving environment because of strong occlusions within the confined space of a vehicle cabin. To overcome these challenges, recent studies leverage the power of deep learning-based approaches, which have shown strong capability in learning discriminative and robust features from visual data. While CNN-based methods only focus on local visual features and unable to effectively handle long-duration videos. Transformer-based models have attracted growing attention due to their ability to capture long-range temporal dependencies and model global contextual relationships within video sequences. One of the most powerful Transformer-based for video representation is proposed by Wang et al. [2], namely VideoMAE, which is developed from ImageMAE model, first designed for static images [3]. The main purpose of the VideoMAE model is to learn effective spatio-temporal representations from videos in a self-supervised manner. To achieve this, a masking strategy is introduced to simulate occlusions, and VideoMAE is trained to reconstruct the missing information from the input videos. This capability makes VideoMAE a promising solution for handling the challenges of DAR in complex in-vehicle environments.

Motivated by the impressive performance of VideoMAE model in HAR [4] and hand gesture recognition (HGR) [5], in this work, we propose a Transformer-based framework for DAR, namely MAE4DAR, where VideoMAE model is employed as a feature extractor to capture the discriminative characteristics of driver actions. Some extensive experiments are conducted on two datasets for driver action recognition that are UTCDriverAct and Driver Action Dataset (DAD) to show the benefit of MAE4DAR framework. The contribution of this work is twofold. First, we introduce a unified pipeline that handles both isolated and continuous recognition scenarios on the same VideoMAE-based backbone. In isolated recognition setting, some experiments are conducted on two datasets for driver action recognition that are UTCDriverAct and Driver Action Dataset (DAD). On UTCDriverAct, MAE4DAR framework achieves strong performance, reaching perfect recognition accuracy in four of the six driver action classes. For DAD, to maintain consistency with our vision-based framework, only RGB images from the body and face views are used. The proposed framework attains recognition accuracies of 80.5% on body view and 82.9% on face view. In continuous setting, the proposed framework yields a frame-wise accuracy of 92.7% on both views, with overlap scores of 0.72 and 0.69 for front-view and rear-view data in UTCDriverAct dataset, respectively. Second, we propose multi-view fusion at score level with three different operators: average, maximum, and multiplication. The main objective of score-level fusion scheme is to combine prediction scores obtained from independently trained single-view models. The experimental results indicate that multi-view fusion consistently improves recognition performance, achieving an overlap score of 0.75 and a frame-wise accuracy of 93.5%, outperforming both single-view settings. Futhermore, an analysis of the trade-off between overlap score and computational cost is conducted for both single-view and multi-view scenarios. The rest of this paper is organized as follows. Some prominent studies on DAR are briefly discussed in Section 2. Next, the proposed framework for continuous driver action recognition is presented in Section3. Some experiments and results are provided and analyzed in Section 4. Finally, several conclusions and future work are mentioned in Section 5.

2. Related Work

In this section, several prominent studies on DAR are reviewed and analyzed to identify their advantages as well as their limitations. To build a discriminative and robust descriptor for action representation, most existing works can be broadly classified into two main approaches including CNN-based and Transformer-based methods. While CNN models mainly capture local features and have difficulty handling long-term dependencies, Transformers are capable of modeling global relationships among elements within a sequence.

2.1. CNN-Based Driver Action Recognition

For the first approach, some early work propose a two-stream network in which a branch is used for visual information and the other branch is employed for motion data. For this, RGB images and their corresponding optical flow images are treated as the input of the two-stream network to extracted independently spatial and temporal information. These extracted features are then concatenated for representing a driver action [6]. In the work is introduced by Tao et al. [6], driver's static attitude and optical flow images are fed forward to two independent CNN-based networks to capture visual and motion information, respectively. Afterward, the feature vectors are weighted and fused to form a final descriptor for characterizing an examined driver

action. However, two main drawbacks of two-stream CNN-based networks are their limited ability to learn long-range dependencies and to model global contextual information. To handle these challenges, some other studies propose to leverage the combination of CNN and RNN models for capturing both spatial and temporal cues [7, 8]. In these studies, CNN models extract spatial features from individual frames, while RNN models capture temporal dependencies between consecutive frames based on CNN-extracted features. Motivated by real-world observations, Behera et al. [7] emphasize the importance of hand-object interation in representing driver activities. Based on this, the authors introduce a multi-stream LSTM-based framework to leverage useful features extracted from the driver posture and hand-object interaction. The integration of CNN and RNN models is also suggested in the study by Hu el al. [8], where a CNN-based network is employed to encode short-term spatial-temporal information and a ConvLSTM model is used to learn long-term temporal dependencies while preserving spatial information. In practice, CNN-RNN architectures focus on modeling temporal dependencies based on spatial features extracted by CNNs, potentially resulting in the loss of motion information. Consequently, some recent work utilize 3D-CNN architectures [9] for learning simultaneously spatial and temporal cues for action representation. In [10], the authors introduce an end-to-end framework for driver action recognition. In this study, both isolated and continuous recognition scenarios are investigated. For this, MoViNet [11] is utilized as a video classifier and then, several post-processing strategies are proposed to localize the staring-time and ending-time for each detected driver activity. Considering the significance of surrounding objects in DAR, such as cell phones, bottles, and cigarettes, several existing studies propose novel frameworks to detect task-relevant objects for identifying driver actions [12]. One of the most impressive studies belonging to this approach is introduced by Tran et al. [13]. In this study, X3D [14], a lightweight 3D-CNN model designed for video understanding, is employed to capture spatio-temporal descriptors for action representation. Although 3D-CNNs achieve high performance in recognition tasks, they still suffer from high computational cost and large memory requirements, which limit their scalability and practical deployment. Therefore, the transition from CNN-based approach to transformer-based approach has become vital in recent studies [15].

2.2. Transformer-Based Driver Action Recognition

In recent years, a large number of studies on DAR have concentrated on proposing a transformer-based framework to build a more effective descriptor for characterizing driver actions. Vision Transformer (ViT) [16] is evaluated as a powerful transformer-based model for image recognition, where an image is considered as a sequence of patches and the global temporal relationships are learned through self-attention mechanism. With the significant assistance of ViT, the recognition performance is improved significantly [17, 18]. Besides ViT model, VideoMAE (Masked Autoencoder) [2, 19] is also a widely used and effective model for video recognition that aims to learn and reconstruct occluded information through a masking strategy. Pizarro et al. [20] introduce a worthwhile framework for DAR by exploring the driver posture information for action recognition. Additionally, in this work, redundant information is removed to speed up the recognition process.

From the above analysis, we can see that the majority of existing work on DAR have just handled single-view data, even in the case of drive action data is collected from multiple views [10, 13]. This is perceived as a research gap and motivation for us to propose a view-aware framework for DAR, in which recognition results for each view are fused to improve the

overall recognition performance. The details of the proposed framework will be presented in the following section.

3. Proposed Method for Continuous Action Recognition

Figure 1 illustrates the overall proposed transformer-based framework for DAR task, which consists of two main phases: training and testing. In the training phase, segmented videos that contain single driver actions are employed to fine-tune a pre-trained VideoMAE model for DAR, while also optimizing the weights of the linear and classification layers. The output of the training phase is a trained model, which is subsequently used in the testing phase for action recognition. In the testing phase, to localize the temporal boundaries for each recognized action, several post-processing strategies are applied on the output of the classification stage. It is worth noting that both single-view and multi-view evaluations are investigated in this work. To this end, several late fusion methods are explored to combine the results obtained from each single-view evaluation. The details of this proposed framework are described below.

3.1. Sampling and Video Patch Embedding

In the first stage, a given input video is first temporally sampled using a sliding window to generate clip-level segments. The primary goal of the sampling step is to allow a deep-learning model to learn local temporal information and improve training efficiency. Let a stream video input be denoted as $V = \{f_1, f_2, \ldots, f_T\}$ where f_i represents the i-th frame. An i-th sample, known as a clip-level segment, is constructed as $C_i = \{f_{i\times s}, f_{i\times s+1}, \ldots, f_{i\times s+w-1}\}$ where w and s are the window size and the temporal stride, respectively, the value of i is in the range of $[0, \frac{T}{s}]$. This means that each clip-level segment contains a total of w frames and is subsequently divided into cube embeddings which serve as input to VideoMAE encoder. In this work, for an arbitrary video, the input frames are resized to a spatial resolution of 224×224. The values of w and s are set to 16 and 15, respectively, and each cube embedding has a size of $16 \times 16 \times 16$.

3.2. VideoMAE Encoder for Action Representation

VideoMAE is known as one of the most effective transformer-based architectures for video representation. In this model, an input video is treated as a sequence of cube embeddings, similar to how sentences are represented as token sequences in natural language processing (NLP). The core idea of VideoMAE is inspired by masked autoencoding, a self-supervised learning paradigm, where a large portion of the input video tokens is randomly masked, and the model is trained to reconstruct the missing spatio-temporal content [19]. As a result, VideoMAE effectively learns robust representations of both appearance and motion dynamics without requiring manual human annotations.

VideoMAE V2 [2] is introduced as an extended version with a more efficient and scalable architecture to learn large-scale video representations. The architecture of VideoMAE v2 is illustrated in Figure 2. Similar to the previous version, VideoMAE V2 uses a high masking ratio of 90% to 95% to create pseudo-occlusion data for the encoder, which is built on a Vision Transformer (ViT) backbone [16] with joint spatial-temporal attention mechanisms. Additionally, the decoding stage adopts a lightweight masking decoder, which reconstructs the masked tokens while reducing computational cost. This design enables efficient pre-training on large-scale action recognition datasets while ensuring strong representation capability. In our framework, we utilize the VideoMAE V2 encoder pretrained on Kinetics-700 dataset as a feature extractor to create a 768-dimensional feature vector for representing a clip-level segment.

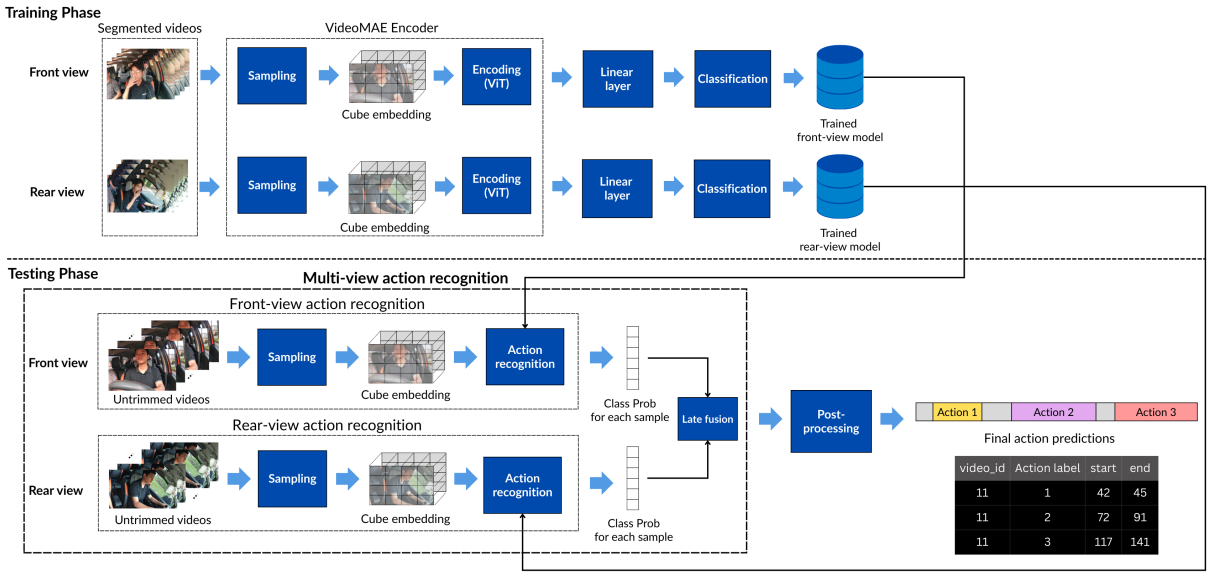


Figure 1. Proposed framework MAE4DAR for driver action recognition.

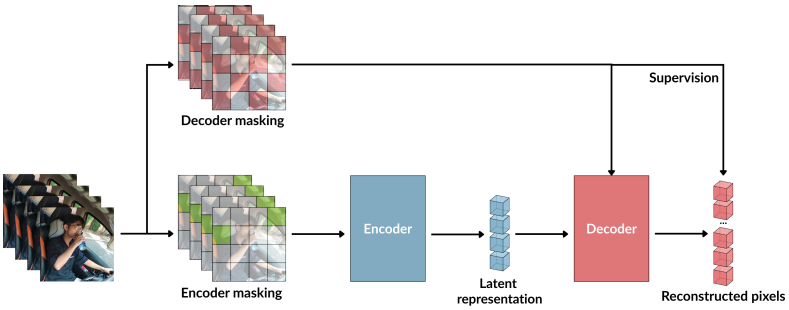


Figure 2. The architecture of VideoMAE model for video understanding task.

3.3. Linear Layer and Classification

The encoded output of the VideoMAE model provides a compact feature representation for each clip-level segment. Specifically, for each sampled clip C_i, the encoder generates a feature vector $\mathbf{z}_i \in \mathbb{R}^{768}$ that captures the spatio-temporal information of the input frames. This feature vector is then fed into a linear classification layer followed by a softmax function to produce class probabilities. Formally, the prediction for the i-th clip from view v is defined as in Equation (1):

$$\mathbf{p}_i^{(v)} = \text{softmax}(W\mathbf{z}_i^{(v)} + b), \tag{1}$$

where W and b denote the learnable parameters of the linear classifier, and $\mathbf{p}_i^{(v)} = [p_{i,1}^{(v)}, p_{i,2}^{(v)}, \ldots, p_{i,K}^{(v)}]$ represents the probability distribution over K action classes for the i-th clip from view $v \in \{\text{front-view}, \text{rear-view}\}$. By sliding the temporal window across the entire video, a sequence of probability vectors is generated, as shown below:

$$\{\mathbf{p}_1^{(v)}, \mathbf{p}_2^{(v)}, \ldots, \mathbf{p}_N^{(v)}\}, \tag{2}$$

where N denotes the total number of sampled clips extracted from the input video. These clip-level predictions serve as the basis for the subsequent temporal localization and multi-view fusion steps.

3.4. Post-Processing and Temporal Action Localization (TAL)

As previously discussed, continuous recognition can be considered as an extension of isolated recognition, where its outputs are integrated to jointly identify actions and determine their temporal boundaries. However, the use of a sliding window mechanism and the limited temporal context in isolated recognition lead to fragmented and discontinuous predictions. To address this issue, a post-processing stage is adopted to improve overall recognition accuracy. The details of several post-processing strategies are described as below. The obtained probability

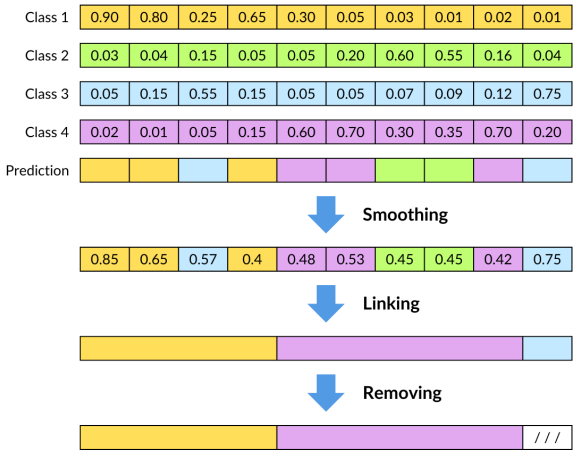


Figure 3. An illustration of several post-processing strategies, including smoothing, linking, and removing.

sequence is further refined through temporal post-processing in order to improve action boundary estimation. These post-processing strategies have been introduced earlier in the work of Zhou et al. [21], including three main substeps: smoothing, linking, and removing. As illustrated in Figure 3, a temporal smoothing operation is first applied using an average filter to reduce short-term fluctuations in prediction scores. After smoothing, action segments are determined by selecting the class with the highest probability for each clip and grouping consecutive clips with the same predicted label into continuous temporal intervals. Adjacent segments with identical labels are then merged to form complete action instances over time.

3.5. Multi-View Score Late Fusion

To exploit complementary visual information from different viewpoints, two independent models are trained for front-view and rear-view videos, respectively. Both models share the same VideoMAE-based architecture, but are optimized separately using view-specific training data. During inference, each model produces a sequence of clip-level probability vectors for the corresponding view. Let $\mathbf{p}_i^{(f)}$ and $\mathbf{p}_i^{(r)}$ denote the prediction vectors from the front-view and rear-view models for the i-th clip, respectively. The final prediction is obtained by combining the two probability vectors using a score aggregation function, presented in Equation (3):

$$\mathbf{p}_i = \mathcal{F}(\mathbf{p}_i^{(f)}, \mathbf{p}_i^{(r)}), \tag{3}$$

where $\mathcal{F}(\cdot)$ denotes a fusion operator. In this work, three different fusion strategies are evaluated, namely the average operator, the maximum operator, and the multiplication operator, as presented in Table 5. The average fusion computes the mean probability scores from the two camera views, assuming that both views contribute equally to the final prediction. Maximum fusion selects the highest probability value across the two views for each class, while multiplication fusion combines the probability scores through element-wise multiplication. These late fusion strategies enable the framework to integrate complementary cues captured from different viewpoints while maintaining a simple and efficient inference pipeline.

4. Experimental Results and Discussion

In this section, we present some extensive experiments on UTCDriverAct to demonstrate the effectiveness of the proposed framework in

Figure 4. Some examples for distracted driver actions captured by two static cameras in front-view and rear-view in UTCDriverAct dataset.

DAR task. Two experimental scenarios are explored in this work, including single-view and multi-view evaluation. All experiments are conducted on a server with the following configuration: an NVIDIA GeForce RTX 4070 GPU with 12 GB VRAM, NVIDIA driver version 550.163.01, and CUDA version 12.4. The operating system is Ubuntu (Linux).

It should be noted that, VideoMAE V2 model with a ViT-L/16 backbone is adopted for action recognition. ViT-L/16 backbone consists of 24 Transformer layers with a patch size of 16×16, and is initialized using pre-trained weights on the Kinetics-700 dataset. VideoMAE V2 is optimized using the Lion optimizer with an initial learning rate of 1×10^{-3}. The training process is performed for 30 epochs with a batch size of 2 using cross-entropy loss function for supervision. These parameter settings are chosen to ensure stable convergence during training while remaining compatible with available computational resources.

4.1. Dataset and Evaluation Metrics

4.1.1. UTCDriverAct Dataset

UTCDriverAct, a driver action recognition dataset, is gathered for the research project funded under grand number B2024-GHA-11. The purpose of UTCDriverAct is to describe several distracted activities of Vietnamese drivers. This dataset is recorded by two static cameras mounted inside the vehicle cabin, capturing driver actions from the front and rear views. The data collection involves 18 volunteers and generates 40 videos, each containing multiple driver actions. To this end, each volunteer is required to perform a sequence of six distracted actions, including *Smoking*, *Texting*, *Calling*, *Drinking*, *Yawning*, and *Talking to Passenger*. Ground-truth annotations are provided in the form of temporal segments with starting-time and ending-time for each driver action appears from any viewpoint. Figure 4 shows some sample videos captured from both front-view and rear-view perspective in UTCDriverAct dataset.

To ensure subject-independent evaluation, the dataset is split according to subject IDs. The first 10 subjects are used for the training phase, while the remaining 8 subjects are utilized for the testing phase. This split mechanism avoids identity overlap between the training and testing sets and reflects a realistic scenario where the model is evaluated on unseen driver actions. The statistic of the training set for each view are summarized in Table 1. As observed in this Table, the training set exhibits class imbalance. To address this issue, an oversampling strategy with a factor of five is applied to the six target driver actions, while the "No Action" class remains unchanged. Additionally, we can see a large variation in terms of average duration for different action classes, from 3.4 seconds for *Yawning* class to 25.1 seconds for *Calling* class. This further highlights the challenges posed by this dataset for DAR.

4.1.2. Driver Activity Dataset

Driver Activity Dataset (DAD) [22] is another publicly available dataset for driver action recognition, which differs from UTCDriverAct dataset in that it provides only trimmed video clips, each corresponding to a single driver action. Some examples for distracted driver actions in Driver Activity Dataset are shown in Figure 5. DAD contains a total of 11 action classes capturing from multiple sensors such as mmWave radars, RGB cameras, and depth cameras, each providing three viewpoints: body, face, and hands. Although this dataset provides multi-modality data, only RGB images from two viewpoints, body and face, are used in this work to ensure consistency with our vision-based framework for driver action recognition. In DAD, *Waiting* and *Driving safety* action classes reflect neutral driving states rather than distracted driving and may introduce bias in performance evaluation. Therefore, these two

Table 1. A statistic about the number of instances and average duration for each class in the training set of UTCDriverAct

Action classes	Number of instances	Average duration (s)
No Action	136	12.7
Smoking	34	4.9
Calling	12	25.1
Texting	12	22.6
Drinking	15	16.0
Yawning	33	3.4
Talking to Passenger	27	17.2

Figure 5. Some examples for distracted driver actions in Driver Activity Dataset [22].

action classes are excluded from our experiments and only the remaining nine driver actions are used for evaluation. It is worth noting that in the original study, this dataset is primarily used for frame-level action recognition with the RGB modality. In this work, we explore the effectiveness of MAE4DAR framework on DAD in isolated recognition scenario.

4.1.3. Evaluation Metrics

For isolated action recognition, accuracy is one of the most widely used metrics. The confusion matrix also provides a detailed analysis of the classification performance between different action classes. It is represented as a square matrix whose size is equal to the number of driver action classes. Each row corresponds to the ground-truth class, while each column corresponds to the predicted class.

For continuous action recognition, two metrics including Frame-wise accuracy and Overlap score are used for performance evaluation. Frame-wise accuracy measures the proportion of correctly classified frames over all frames in a continuous video sequence. The formulation of this metric is presented in Equation 4, as follows:

$$\text{Frame-wise Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}. \quad (4)$$

Where N_{correct} denotes the number of correctly predicted frames and N_{total} is the total number of examined frames. Although frame-wise accuracy provides an intuitive measure of classification performance, it is sensitive to class imbalance.

Overlap score is computed based on a pre-defined temporal threshold, called δ. Let a ground-truth action segment g have start time g_s and end time g_e, and a predicted action segment p with the same action label have start time p_s and end time p_e. A prediction is considered a match

if $p_s \in [g_s - 2s, g_s + 2s]$ and $p_e \in [g_e - 2s, g_e + 2s]$. The overlap score between a matched prediction and it's ground-truth is defined as in the following Equation:

$$os(p, g) = \frac{\max\left(\min(g_e, p_e) - \max(g_s, p_s), 0\right)}{\max(g_e, p_e) - \min(g_s, p_s)} \quad (5)$$

A predicted segment which do not temporally overlap with any ground-truth one is assigned an overlap score of zero. The final overlap score is obtained by averaging the overlap scores across all predicted and ground-truth action segments.

4.2. Results and Discussions

4.2.1. Performance Evaluation on Isolated Action Recognition

It is worth noting that isolated recognition operates on segmented videos which consist of single driver action and action recognition is considered as classification task. Table 2 presents isolated action recognition performance on two examined datasets with different camera views. For UTCDriverAct dataset, the proposed method achieves high accuracy, reaching 94.9% on the front view and 97.9% on the rear view, while the performance on DAD is lower, with 80.5% on the body view and 82.9% on the face view. This gap suggests that DAD is more challenging due to the larger number of action classes and the higher inter-class similarity.

Table 2. Isolated action recognition accuracy in UTCDriverAct and Driver Activity Dataset (%)

Dataset	**View**	**Accuracy**
UTCDriverAct	Front view	94.9
	Rear view	97.9
Driver Activity Dataset	Body view	80.5
	Face view	82.9

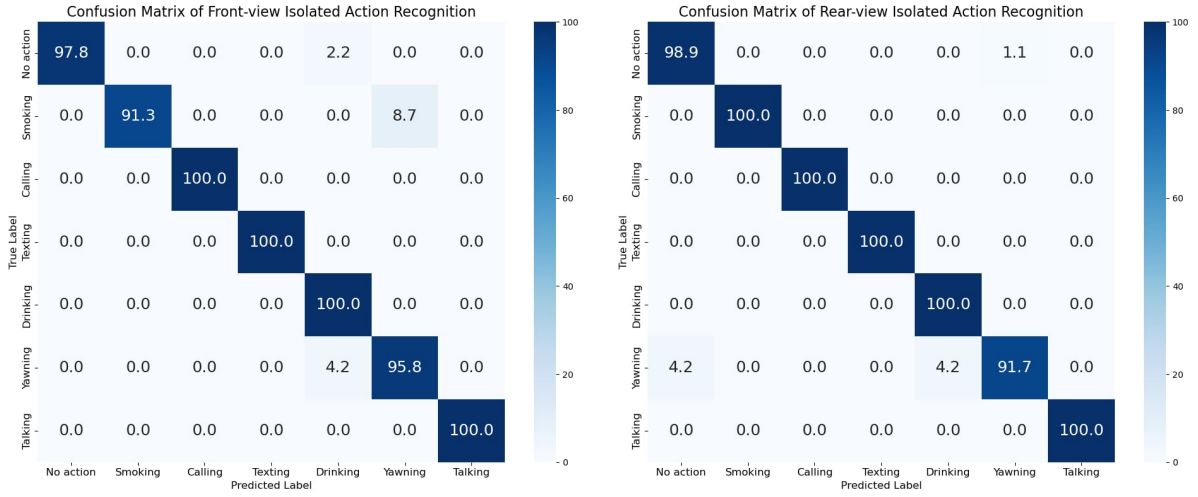
Table 3 and Figure 6 demonstrate the recognition accuracy obtained for isolated recognition scenario on each camera view of the UTCDriverAct dataset. We can observe that recognition performance achieves impressive results for all action classes, even obtaining the perfect accuracy on four of six driver action classes for both front and rear views, including *Calling*, *Texting*, *Drinking*, and *Talking to Passenger*. For *Smoking* and *Yawning* classes, the recognition accuracies are lower, with values of 91.3% and 95.8% on front-view data and 100% and 91.7% on rear-view data, respectively. These results reflect a realistic scenario that *Smoking* and *Yawning* activities are more challenging than other driver activities in action recognition.

4.2.2. Performance Evaluation on Continuous Action Recognition

For continuous action recognition, frame-wise accuracy and temporal overlap score are evaluated. It is worth noting that both single-view and multi-view data are used for performance evaluation. The obtained results in this experimental scenario are shown in Table 4. By observing this Table, the values of frame-wise accuracy and overlap score on front-view data are 92.7% and 0.72, respectively. These evaluation metrics on rear-view data are 92.7% and 0.69, respectively. From these results we declare that the proposed framework can yield a higher performance with front-view data, the overlap score on achieves an improvement of 0.03 compared to that on rear-view data. For multi-view data, a late fusion relied on mean operator is used to integrate the recognition results on both single-view data. In this case, the proposed framework achieves 93.5% and 0.75 in terms of frame-wise accuracy and overlap score, respectively. The improvement in both terms of evaluation metrics indicate that by leverage information captured from different camera views, the recognition accuracy can be increased. However, this late fusion mechanism may lead to a higher computational cost for action recognition task. Additionally, inference times for both single-view and multi-view settings are provided in Table 2. Noted that these values are

Table 3. Recognition accuracy on isolated videos in UTCDriverAct (%)

Action classes	Front-view data	Rear-view data
No Action	97.8	98.9
Smoking	91.3	100.0
Calling	100.0	100.0
Texting	100.0	100.0
Drinking	100.0	100.0
Yawning	95.8	91.7
Talking to Passenger	100.0	100.0
All classes	94.9	97.9



(a) (b)

Figure 6. Confusion matrices for isolated action recognition scenario: a) on front-view data and b) on rear-view data.

Table 4. Performance evaluation on both single-view data and multi-view data for continuous action recognition

Evaluation data	Frame-wise Accuracy	Overlap Score	Inference time (s)
Front-view	92.7	0.72	36.6
Rear-view	92.7	0.69	36.3
Multi-view	**93.5**	**0.75**	71.9

computed on the entire testing set for each single viewpoint and the multi-view data. It clear that handling multi-view data requires approximately twice the inference time compared to single-view data. This issue can be addressed by employing a parallel mechanism during the computation process.

Figure 7 illustrates the class-wise overlap score performance on the front-view, rear-view, and the multi-view data. Overall, the results demonstrate that combining predictions from both views consistently improves the localization performance across most action classes.

From this Figure, the proposed framework can reach a higher performance on rear-view data for several actions, such as *Smoking*, *Drinking*,

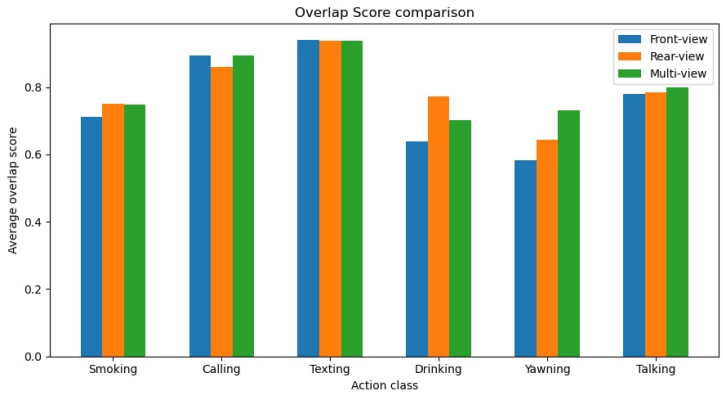


Figure 7. Class-wise overlap score comparison on single-view and multi-view data.

Table 5. Comparison of different multi-view fusion strategies

Fusion strategies	**Average overlap score**
Average operator	**0.75**
Max operator	**0.75**
Multiplication operator	0.73

and *Yawning*. These actions typically involve noticeable hand or head movements that are more clearly captured from the rear camera perspective. In contrast, the proposed framework performs better on front-view data for *Calling* class, where facial orientation and upper-body posture provide useful visual cues. For *Texting* class, the proposed framework achieves relative results, suggesting that this action has distinctive motion information that can be reliably detected from both viewpoints. For multi-view data, the obtained results once again confirm that by exploiting useful information from different camera viewpoints, the recognition performance can be significantly improved. Overlap score is increased on almost driver action classes when deal with multi-view data. This indicates that combining complementary information from both viewpoints helps reduce ambiguities in temporal localization.

Furthermore, to analyze the effectiveness of different late fusion strategies for multi-view action recognition, three commonly used fusion operators including average, maximum fusion, and multiplication are investigated. The performance comparison of these fusion strategies is presented in Table 5. The results show that both the average operator and the max operator achieve the best performance with an overlap score of 0.75, indicating that simple score-level fusion can effectively combine complementary information from the two camera viewpoints. In contrast, the multiplication operator yields slightly lower performance, suggesting that suppressing predictions when one view has lower confidence may negatively affect the final recognition accuracy. In addition to score-level fusion, experiments with feature-level fusion are also conducted, where feature representations are first extracted from different views using fine-tuned models and then combined for classification. Table 6 provides a comparison of recognition accuracy when exploiting different fusion schemes. From this Table, we can see that feature-level fusion scheme achieves an overlap score of 0.73 using either average operator or maximum operator, which is 0.02 lower than that of the score-level fusion scheme. This reduction can be explained by each classifier being trained independently on single-view data and not being optimized for a feature-level fusion strategy. In contrast, score-level fusion operates on prediction

Table 6. Comparison of feature-level and score-level multi-view fusion strategies

Fusion strategies	**Overlap score**
Feature-level fusion with average pooling	0.73
Feature-level fusion with max operator	0.73
Proposed score-level fusion	**0.75**

probabilities, enabling each single-view model to maintain robust performance. These results once again confirm the effectiveness of MAE4DAR for driver action recognition.

5. Conclusions

In this paper, we introduce a Transformer-based framework for DAR, where VideoMAE V2 is utilized as a video encoder to extract informative cues for driver action representation. Some extensive experiments are performed on UTCDriverAct and DAD datasets to prove the effectiveness of MAE4DAR framework in DAR task. Firstly, we evaluate the performance of MAE4DAR on both isolated and continuous settings. For isolated recognition, MAE4DAR achieves accuracy of 80.5% (body view) and 82.9% (face view) on DAD, and 94.9% (front view) and 97.9% (rear view) on UTCDriverAct. In the continuous recognition setting, the proposed framework obtains a frame-wise accuracy of 92.7% on both views, with overlap scores of 0.72 and 0.69 on front-view and rear-view data, respectively. Secondly, we propose leveraging a late fusion scheme based on prediction scores derived from single-view data to further enhance overall recognition performance in DAR. By applying score-level fusion using either the average or maximum operator, the proposed framework achieves a frame-wise accuracy of 93.5% and an overlap score of 0.75. These results highlight the benefit of exploiting complementary information from different viewpoints, especically in scenarios involving severe occlusion or missing views. Furthermore, a comparison on inference time when dealing with single-view data and multi-view data is conducted to indicate the trade-off between recognition accuracy and computational cost. In future work, we will explore adaptive fusion strategies and end-to-end multi-view learning to further enhance recognition robustness.

Acknowledgement

This research was funded by the Vietnam Ministry of Education and Training under grant number B2024-GHA-11.

References

[1] B. Zhang, J. Wang, J. Fu, J. Xia, Driver Action Recognition Using Federated Learning, in: Proceedings of the 7th International Conference on Communication and Information Processing, 2021, pp. 74–77. DOI: https://doi.org/10.1145/3507971.3507985.

[2] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, Y. Qiao, Videomae V2: Scaling Video Masked Autoencoders with Dual Masking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14549–14560. DOI: https://doi.org/10.48550/arXiv.2303.16727.

[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked Autoencoders Are Scalable Vision Learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16000–16009. DOI: 10.1109/CVPR52688.2022.01553.

[4] H. Yin, R. O. Sinnott, G. T. Jayaputera, A Survey of Video-Based Human Action Recognition in Team Sports, Artificial intelligence review, Vol. 57, No. 11, 2024, pp. 293. DOI: https://doi.org/10.1007/s10462-024-10934-9.

[5] P.-D. Nguyen, V.-T. Tran, D.-K. Ngo, V.-D. Le, H.-A. Nguyen, T.-B. Nguyen, H.-Q. Nguyen, T.-L. Le, CoboGesture: A Continuous Hand Gesture Dataset

and Recognition Method for Human-Collaborative Robot Interaction, IEEE Access (2026). DOI: 10.1109/ACCESS.2026.3664027.

[6] C. Tao, S. Ma, Driver Distraction Recognition with Pose-Aware Two-Stream Convolutional Neural Network, in: Eighth International Conference on Electromechanical Control Technology and Transportation (ICECTT 2023), Vol. 12790, SPIE, 2023, pp. 1350–1359. DOI: https://doi.org/10.1117/12.2689437.

[7] A. Behera, A. Keidel, B. Debnath, Context-Driven Multi-Stream LSTM (M-LSTM) for Recognizing Fine-Grained Activity of Drivers, in: German Conference on Pattern Recognition, Springer, 2018, pp. 298–314. DOI: https://doi.org/10.1007/978-3-030-12939-2_21.

[8] Y. Hu, M. Lu, C. Xie, X. Lu, Video-Based Driver Action Recognition via Hybrid Spatial-Temporal Deep Learning Framework, Multimedia systems, Vol. 27, No. 3, 2021, pp. 483–501. DOI: https://doi.org/10.1007/s00530-020-00724-y.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497. DOI: https://doi.org/10.1109/ICCV.2015.510.

[10] H.-Q. Nguyen, T.-B. Nguyen, T. K. Tran, V.-N. Hoang, T.-L. Le, T.-H. Tran, H. Vu, End-to-End Deep Learning-Based Framework for Driver Action Recognition, in: 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), IEEE, 2022, pp. 1–6. DOI: https://doi.org/10.1109/MAPR56351.2022.9924944.

[11] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, B. Gong, MoVinets: Mobile Video Networks for Efficient Video Recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 16020–16030. DOI: 10.1109/CVPR46437.2021.01576.

[12] H. I. Qamar, U. Saeed, M. Hussain, Driver Distraction Detection Using A Multi-Stream Deep Fusion Network, in: International Conference on Computing & Emerging Technologies, Springer, 2023, pp. 97–104. DOI: https://doi.org/10.1007/978-3-031-77617-5_9.

[13] M. T. Tran, M. Q. Vu, N. D. Hoang, K.-H. N. Bui, An Effective Temporal Localization Method with Multi-View 3D Action Recognition for Untrimmed Naturalistic Driving Videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3168–3173. DOI: https://doi.org/10.1109/CVPRW56347.2022.00357.

[14] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional Two-Stream Network Fusion for Video Action Recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1933–1941. DOI: https://doi.org/10.1109/CVPR.2016.213.

[15] H.-H. Bui, K.-H. Bui, T.-L. Vo, H.-Q. Nguyen, T.-B. Nguyen, T.-T. Dao, T.-L. Le, HIT4DAR: Holistic Interaction Transformer for Driver Action Recognition, Transport and Communications Science Journal, Vol. 77, No. 4, 2026, pp. 500–514. DOI: https://doi.org/10.47869/tcsj.77.4.12.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021). arXiv:2010.11929, DOI: https://arxiv.org/abs/2010.11929.

[17] Y. Ma, L. Yuan, A. Abdelraouf, K. Han, R. Gupta, Z. Li, 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, 2023, pp. 5287–5294. DOI: https://doi.org/10.1109/CVPRW59228.2023.00557.

[18] N. Sengar, I. Kumari, J. Lee, D. Har, PoseViNet: Distracted Driver Action Recognition Framework Using Multi-View Pose Estimation and Vision Transformer, arXiv preprint arXiv:2312.14577 (2023). DOI: https://doi.org/10.48550/arXiv.2312.14577.

[19] Z. Tong, Y. Song, J. Wang, L. Wang, VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training, Advances in neural information processing systems, Vol. 35, 2022, pp. 10078–10093. DOI: https://doi.org/10.52202/068431-0732.

[20] R. Pizarro, R. Valle, L. M. Bergasa, J. M. Buenaposada, L. Baumela, Pose-Guided Multi-Task Video Transformer for Driver Action Recognition, arXiv preprint arXiv:2407.13750 (2024). DOI: https://doi.org/10.48550/arXiv.2407.13750.

[21] W. Zhou, Y. Qian, Z. Jie, L. Ma, Multi View Action Recognition for Distracted Driver Behavior Localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5375–5380. DOI: https://doi.org/10.1109/CVPRW59228.2023.00567.

[22] G.-H. Li, H.-C. Chiang, Y.-C. Li, S. Shirmohammadi, C.-H. Hsu, A Driver Activity Dataset with Multiple RGB-D Cameras and mmwave Radars, in: Proceedings of the 15th ACM Multimedia Systems Conference, 2024, pp. 360–366. DOI: https://doi.org/10.1145/3625468.3652181.