Transport and Communications Science Journal

HIT4DAR: HOLISTIC INTERACTION TRANSFORMER FOR DRIVER ACTION RECOGNITION

Hoang Hiep Bui[1], Khanh Huyen Bui[1], Thien Linh Vo[2], Hong Quan Nguyen[3], Thuy Binh Nguyen[2*], Thanh Toan Dao[2], Thi Lan Le[1]

[1]SigM Lab, School of Electrical and Electronic Engineering (SEEE), Hanoi University of Science and Technology, Hanoi, Vietnam

[2]Universiy of Transport and Communications, Hanoi, Vietnam

[3]Viet-Hung Industrial University, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 20/02/2026
Revised: 22/03/2026
Accepted: 23/03/2026
Published online: 15/05/2026
https://doi.org/10.47869/tcsj.77.4.12

*Corresponding author
Email: thuybinh_ktdt@utc.edu.vn

Abstract. Traffic accidents remain a critical issue worldwide, especially in the context of the rapidly growing number of vehicles on the road. This leads to the need for an automatic system to recognize distracted driver actions and provide early warnings to enhance road safety. In fact, driver action recognition (DAR) can be considered as a subfield of human action recognition (HAR). In addition to the common challenges of HAR, DAR must address additional difficulties, including fine-grained and subtle hand movements, self-occlusion, and complex interactions with multiple objects. To overcome these challenges, we leverage Holistic Interaction Transformer (HIT) network, originally designed for HAR, to recognize driver activities from video sequence. The proposed method named HIT4DAR (Holistic Interaction Transformer for Driver Action Recognition). Some experiments are conducted on UTCDriverAct to show the effectiveness of HIT network in DAR task. The overall performance across the six actions of interest achieves a Video-mAP of 47.8%. In particular, the recognition accuracy for *Texting* activity reaches 78.9%. Furthermore, an ablation study is performed to investigate the influence of pose estimation models on recognition accuracy. The experimental results indicate the trade-off between recognition accuracy and computational efficiency in different pose estimation models. This comprehensive analysis provides useful recommendations for the research community when deploying the proposed framework in real-world DAR systems.

Keywords: Driver Action Recognition, Transformer-based Action Recognition, Human-Object Interaction.

1. INTRODUCTION

In recent years, driver activity recognition has become a promising research direction in intelligent transportation systems. This problem involves recognizing and analyzing drivers' behaviors while participating in traffic. This is considered as one of the critical task in an Advanced Driver Assistance System (ADAS). From this, ADAS can provide necessary warnings for the driver do reduce the occurrence of unfortunate traffic accidents. DAR is treated as a branch of human activity recognition (HAR). Consequently, it shares common challenges with general HAR, such as background clutter, intra-class variability, and inter-class similarity. Furthermore, DAR also introduces additional difficulties, including dynamic backgrounds caused by vehicle motion and strong occlusions inside the vehicle cabin [1,2]. To overcome the aforementioned challenges in DAR, most existing studies exploit informative cues obtained from either wearable sensors or camera-based systems. However, the wearable-based approach might cause discomfort to users because these sensors must be attached to the human body. Additionally, the performance of wearable-based systems is highly sensitive to sensor placement and orientation. Even minor shifts during driving can lead to significant reduction in recognition accuracy [3]. For these reasons, many recent studies have focused on the use of vision sensors (cameras) for data acquisition. In fact, vision data can provide richer and more detailed information through capturing body posture, motion trajectories, and interactions with objects or environment. These are the necessary factors to achieve higher performance in DAR task [4].

Additionally, DAR task is addressed in two scenarios that are isolated recognition and continuous recognition. Isolated recognition aims at determining the name of activity class for a segmented video, each corresponding to a single human action, and is typically formulated as a classification task. In contrast, continuous recognition analyzes untrimmed video sequences that contain multiple consecutive actions. Continuous recognition not only determines the action class but also estimates the temporal boundaries of each action, including starting- and ending-times [5]. Compared with isolated recognition, continuous recognition presents greater challenges; consequently, its performance is generally lower. In practice, continuous recognition can be regarded as an extension of isolated recognition, in which classification results are further refined through a post-processing stage to determine the temporal boundaries of each predicted action.

To address HAR problem, most current research have concentrated on two critical approaches that are CNN-based (Convolutional Neural Network) and Transformer-based. The crucial target of CNN-based methods is to extract both temporal and spatial information for action presentation [6,7,8]. Although achieving some important milestones, CNN-based methods still can not deal with temporal relationship with the variation in video sequence length. Fortunately, this difficulty can be overcome by utilizing a transformer architecture with the ability of encoding and decoding information [9,10,11,12]. Among them, Holistic Interaction Transformer Network (HIT), introduced in the work of Faure et al. [13], concentrates on exploiting essential information including hand and human pose. HIT network is considered as a comprehensive bi-modal framework that consists of an RGB stream and a pose stream. With the assistance of the two-stream model, HIT network has focused not only on person, object, and hand but also on their interactions, enable improving performance for HAR task. Based on the observation that most driver actions involve interactions between the hands and surrounding objects (e.g., a mobile phone, water bottle, or cigarette), we leverage the potential of the HIT network for the driver action recognition task in this work. The proposed

method is named HIT4DAR (Holistic Interaction Transformer for Driver Action Recognition). A remarkable point of HIT4DAR is its ability to exploit informative cues from interactions among human body, hands, and task-relevant objects for action representation. As above mentioned, HIT model is originally designed for HAR, where most human keypoints can be reliably estimated when the whole human body is clearly visible. Meanwhile, HIT4DAR is developed for the DAR task to address the key challenges caused by severe occlusions in confined environments such as vehicle cabins.

Experiments have been conducted on UTCDriverAct to show the effectiveness of HIT4DAR. The overall performance across the six actions of interest achieves a Video-mAP of 47.8%. In particular, the recognition accuracy for the *Texting* activity reaches 78.9%. Furthermore, an ablation study is performed to investigate the influence of pose estimation models on recognition accuracy. The experimental results indicate the trade-off between recognition accuracy and computational efficiency across different pose estimation models. This comprehensive analysis provides useful recommendations for the research community when deploying the proposed framework in real-world DAR systems.

2. RELATED WORKS

As aforementioned, most existing studies on driver action recognition task are categorized into two main approaches: CNN-based and Transformation-based methods. Furthermore, CNN-based methods are commonly classified into three categories: two-stream CNN-based models, CNN-RNN hybrid architectures, and 3D CNN frameworks. The main target of these methods is to build an effective spatio-temporal description for action representation. Unlike CNNs and RNNs, Transformers use self-attention mechanism to model global interactions among all elements in a sequence. In this section, some remarkable studies on driver action recognition are briefly discussed to highlight their advantages as well as their limitations. Based on this, we propose to leverage the HIT network, originally designed for HAR, to address several challenges in the DAR task.

2.1. CNN-based action recognition

In the first branch, visual and motion information are processed independently in the two-stream network and then, they are combined for human action representation. In the study of Simonyan et al. [14], RGB images and their optical flow are treated as input of each CNN network to extract corresponding appearance and motion features. However, a major drawback of this framework lies in its inability to explicitly determine which human body parts are moving and to localize them across consecutive frames. To handle this limitation, Feichtenhofer et al. [15] propose to adopt ResNet model [16] into two-stream framework. By incorporating residual connections, ResNet enables the effective training of deep architectures, leading to robust feature representation learning and significantly improved performance in HAR. In [5], the authors introduce an end-to-end framework for driver continuous recognition. In this work, MoViNet [17] is employed as a video classifier that performs both action representation learning and classification tasks. Then, several postprocessing strategies are proposed to provide the final results for continuous recognition. MoViNet is a lightweight CNN architecture specifically designed for efficient video understanding. Its compact design and low computational cost make it highly suitable for deploying recognition systems on hardware platforms, especially resource-constrained edge devices. In practice, CNNs are mainly designed to capture local spatial and short-term temporal cues, which limits their ability to model long-term dependencies required for recognizing complex activities over extended sequences.

To overcome the above limitation, some other studies have concentrated on the second categorize, CNN-RNN hybrid models [6]. While CNNs are designed to extract appearance information, RNNs aim to capture temporal dependencies among features extracted from consecutive frames. By modeling these temporal relationships, only discriminative information is retained for action representation, leading to improved performance in HAR tasks. However, these studies primarily focus on isolated recognition, in which each segmented video contains only a single human action. Such a setting is not suitable for realistic HAR systems, where multiple human actions occur continuously. Related to continuous recognition, some current works emphasize the importance of task-relevant objects (e.g., phones, cigarettes, and bottles) in action representation [7, 8]. In [7], the authors introduce a novel framework which incorporates deformable convolution and dilated convolution to detect small-sized objects such as cigarettes and mobile phones, which are essential for accurate activity recognition. In the work of Qamar et al. [8], Multi-stream Deep Fusion Network (MDFN) is proposed for modeling driver pose and human-object interactions to enhance recognition performance. However, this combination of multi-stream CNNs requires a high computational cost and it is difficult to deploy this framework in a realistic system.

Another CNN-based approach for action recognition is the use of 3D-CNN networks. The advantage of 3D-CNNs compared to 2D-CNNs lies in their ability to simultaneously learn spatial and temporal information, enabling effective capture of motion information over consecutive frames. This capability allows 3D-CNNs to represent spatio-temporal dynamics in an end-to-end manner, making them more suitable for action recognition task. One of impressive studies belonging to this approach is introduced by Tran et al. [18], which focuses on continuous action recognition. In this work, X3D [19] is used to extract spatio-temporal features for action representation, and then some post-processing strategies are performed to identify the occurring action and its temporal occurrence. Although achieving a higher performance in recognition task, 3D-CNNs suffer from high computational cost and large memory requirement, which significantly limit their scalability and practical deployment. Moreover, they primarily capture short-term temporal constraints and struggle to model long-term dependencies in continuous action recognition.

2.2. Transformer-based Driver Action Recognition

For transformer-based approach, VideoMAE [20], an extension of ImageMAE (Masked Autoencoder) - first designed for static images, is proposed as one of the most powerful models for HAR. The main purpose of VideoMAE is to enable the model to learn and reconstruct occluded information generated by a masking strategy. Inspired by the power of VideoMAE model information construction, Pizarro et al. [10] propose an effective framework which integrates pose estimation and action recognition into a unified deep learning architecture. Additionally, this framework attempts to remove tokens corresponding to redundant information to speed up the recognition process.

Besides VideoMAE, Vision Transformer (ViT) [21] is evaluated as one of the most effective models for video understanding tasks. It is worth noting that ViT is initially introduced for Natural Language Processing (NLP) and when applied to computer vision, it models an image as a sequence of patches. Several recent works have explored the potential of ViT to enhance the recognition performance in driver action recognition [9, 11, 12]. In these works, ViT model is employed to build a more discriminative and robust descriptor for action representation, which helps to achieve a higher recognition accuracy. Nevertheless, this approach is primarily limited by its substantial computational cost and significant memory

requirements. To overcome these challenges, Xu et al. [22] introduce spatio-temporal decoupling attention transformer framework (SDA-TR) which is based on skeleton data for DAR task. In this work, the authors observe that several popular pose estimation methods are not well suited to narrow spaces such as a vehicle cabin. Consequently, only most informative joints are selected for representing an examined action to reduce computational cost as well as storage memory.

From the above analysis, we realize that most of existing work have just focused on isolated recognition, which is treated as classification task. Although this approach can achieve some remarkable results, it does not accurately reflect realistic scenarios in which human actions occur continuously. This is the motivation for us to explore some challenges in continuous recognition approach. Further details of the proposed framework are presented in the next section.

3. OVERALL PROPOSED FRAMEWORK

Inspired by the impressive results in the study of Faure et al. [13], we propose to adopt HIT network for driver action recognition, named HIT4DAR (Holistic Interaction Transformer for Driver Action Recognition). Figure 1 illustrates the overall pipeline of the proposed framework, consisting of four main modules: Human/Object detection and pose estimation, Human-Object interaction, Attentive fusion, and Temporal interaction. Firstly, a given video stream is forwarded to Human/Object detection and pose estimation module to determine the regions of interest and extract skeleton data, respectively. Secondly, Human-Object interaction module to capture the interactions among drivers, objects, and their hands, resulting in more discriminative and effective action representations. Then, these two streams are combined through Attentive fusion stage and Temporal interaction focuses on learning temporal relationships across consecutive frames to identify which action is occurring.

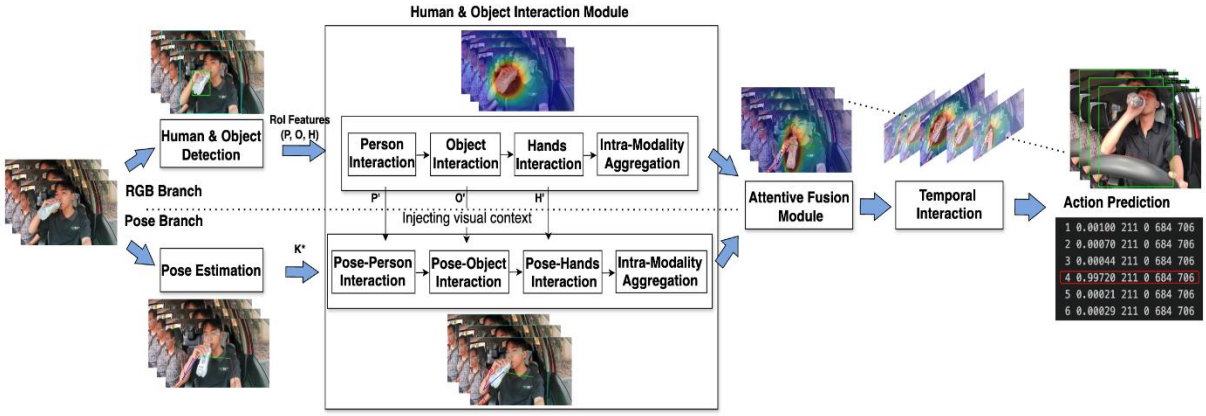


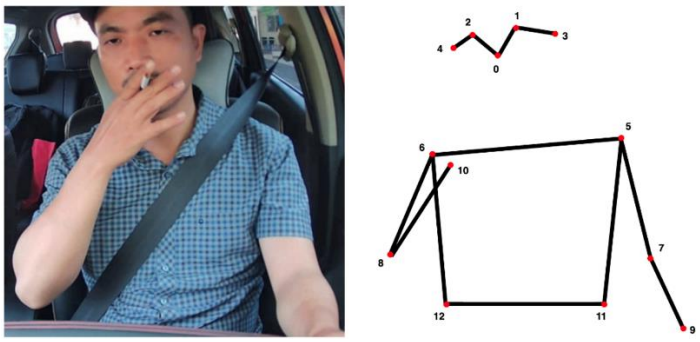
Figure 1. Overview of the proposed framework HIT4DAR for driver action recognition.

It should be noted that the proposed method is built upon a bi-modal framework with separate RGB and pose streams. The primary target of each stream is to learn the interaction between the target person and surrounding objects. Furthermore, Intra-Modality Aggregation (IMA) module is integrated into HIT network to filter and aggregate information from each interaction module. This strategy is applied on both RGB and pose branches to learn informative interaction cues while to compress redundant information. The outputs of the two

streams are fused via an attentive fusion module, and temporal dependencies are exploited to recognize the action. By exploiting the complementarity between visual information and skeleton data, the proposed framework can achieve robustness against typical challenges in driving environments, including illumination variations and partial occlusions. The details of these modules are described below.

3.1. Region of Interest (RoI) localization and feature extraction

At this stage, the input stream is fed into both RGB and pose branches to capture highly discriminative and robust features for action representation. In RGB branch, two sub-tasks are simultaneously performed: video representation and regions of interest (RoIs) localization. To accomplish these two sub-tasks, SlowFast model [23] and Faster R-CNN [24] are employed for video representation and RoIs localization, respectively. SlowFast model consists of two parallel pathways: the slow pathway focuses on spatial semantics at a low frame rate, while the fast pathway emphasizes motion dynamics at a high frame rate. This architecture enables SlowFast to learn valuable spatio-temporal features for video representation. For RoIs localization, the proposed framework aims to define the bounding boxes for the driver, their hands, and some relevant-task objects, such as bottles, phones, and cigarettes. Then, RoIAlign method [25] is utilized to align extracted video features for the driver ($\mathcal{P}$), hands ($\mathcal{H}$), and objects ($\mathcal{O}$). In pose branch, the pre-trained model ResNet-50 [16] is employed as backbone of pose estimation method to encode the driver pose. It is worth noting that due to the limited space within the vehicle cabin, only 13 keypoints are considered in this study. All investigated driver actions primarily involve hand-object interactions, consequently, these 13 keypoints provide sufficient information for effective action representation. Figure 2 illustrates extracted keypoints for human pose in DAR task.



Index	Key point
0	Nose
1	Left eye
2	Right eye
3	Left ear
4	Right ear
5	Left shoulder
6	Right shoulder
7	Left elbow
8	Right elbow
9	Left wrist
10	Right wrist
11	Left hip
12	Right hip

Figure 2. An example of 13 keypoints for human pose in HIT4DAR framework.

3.2. Interaction Modelling in the RGB branch

As illustrated in Figure 1, the RGB branch consists of three crucial components that are person interaction, object interaction, and hands interaction. While person interaction module captures the interaction between persons within the current frame, object and hands interaction attempt to learn the person-object and person-hands interaction, respectively. For this task, HIT network leverages the power of cross-attention mechanism where the target person feature is

defined as the query, and the key and the value are acquired from the object and hand features. The formulation of this cross-attention mechanism is as shown in [13]:

$$A(i) = \text{softmax}\left(\frac{Q_i \times K_i^T}{\sqrt{d_v}}\right) \times V_i \quad i \in (\mathcal{P}, \mathcal{O}, \mathcal{H}) \tag{1}$$

Where d_v is the visual embedding dimension, Q_i, K_i and V_i are the query, the key, and the value matrix, respectively. It is worth noting that cross-attention mechanism takes only person features when computing person interaction $A(\mathcal{P})$ computation. However, for person-object and person-hand interaction, the output of each interaction block and object/hand features are treated as the input.

3.3. Interaction Modeling in the pose branch

Like the RGB branch, cross-attention mechanism is also used to learn the pose-person, pose-object, and pose-hands interactions. Firstly, pose features are extracted from pose data by exploiting light transformer encoder, denoted as $\mathcal{K}^*$. It is worth noting that $\mathcal{P}'$, $\mathcal{O}'$, and $\mathcal{H}'$ are the output of $A(\mathcal{P})$, $A(\mathcal{O})$, and $A(\mathcal{H})$ respectively. For this, HIT network computes $A(\mathcal{K}^*, \mathcal{P}')$, $A(\mathcal{O}')$, $A(\mathcal{H}')$ based on cross-attention mechanism for pose-person, pose-object, and pose-hands respectively. The following formulation is presented in the following equation [13]:

$$A(\mathcal{K}^*, \mathcal{P}') = \text{softmax}\left(\frac{Q' \times K'^T}{\sqrt{d_p}}\right) \times V' \tag{2}$$

Where, d_p is the dimension of pose features, Q' and K' are defined as the output of a transformation applied on $\mathcal{K}^*$ and $\mathcal{P}'$, respectively, V' is known as the value matrix. As shown in the above equation, pose interaction module integrates visual information in pose interaction computation. This strategy allows HIT network to leverage visual appearance cues for effectively distinguishing complex driver poses.

Figure 3. An example for interaction modeling in both RGB and pose branches (best view in color).

The interaction modeling in both RGB and pose branches are illustrated in Figure 3. As seen in this Figure, the interaction between person, hand, and object is focused in the RGB image. While in the pose branch, HIT model attempts to extract informative joints to represent human-body geometric structure and capture the distinctive motion characteristics of human actions. This effectively guides the model's attention toward the arm, elbow, and wrist keypoints that are critical for action recognition.

3.4. Attentive Fusion and Temporal Interaction Modules

To leverage both visual information and skeletal data, an Attentive Fusion Module (AMF) is integrated into the HIT network to fuse these two feature representations. Unlike traditional concatenation techniques, AFM utilizes the self-attention mechanism to dynamically evaluate the contribution of each modality. From this, AFM assigns higher weights to the stream that provides more reliable information. Furthermore, these fused features are passed through a Temporal Interaction Unit to incorporate long-term temporal information for action representation. The output of Temporal Interaction Unit is known as a probability distribution across all action classes for a given input video. With the great assistance of these two modules, action recognition performance is improved significantly.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Dataset and measurement metrics

4.1.1. UCTDriverAct dataset for DAR

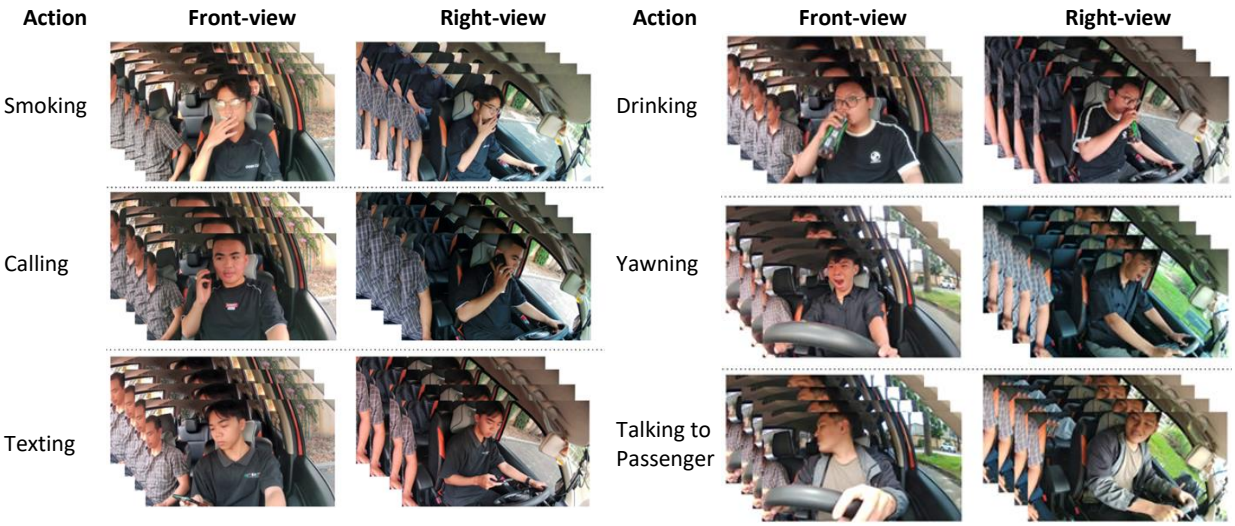


Figure 4. Some examples of UTCDriverAct dataset.

UTCDriverAct dataset is collected as part of research project funded under grand number B2024-GHA-11 for driver action recognition. The main objective of gathering this dataset is to explore distinctive characteristics and behavioral habits of drivers in Vietnam. This dataset consists of total 40 videos collected from two static synchronized cameras of 18 volunteers. Each of the volunteers perform six examined actions sequentially, including *Smoking*, *Drinking*, *Calling*, *Texting*, *Yawning*, and *Talking to Passenger*. All video sequences are captured at a frame rate of 30 frames per second (fps). The duration of each video sequence is approximately four to five minutes. It is worth noting that two static cameras are mounted at two different positions: front-view and rear-view, to provide better observation for driver actions. Figure 4 shows some examples of six distracted driver actions. Among 40 videos of UTCDriverAct, 24 videos correponding to the first 10 IDs are used for the training phase and the remaining (16 videos of the last 8 IDs) are employed for the testing phase. Detailed information about this dataset is available at the following link: https://github.com/vothienlinh/driver-actions.git.

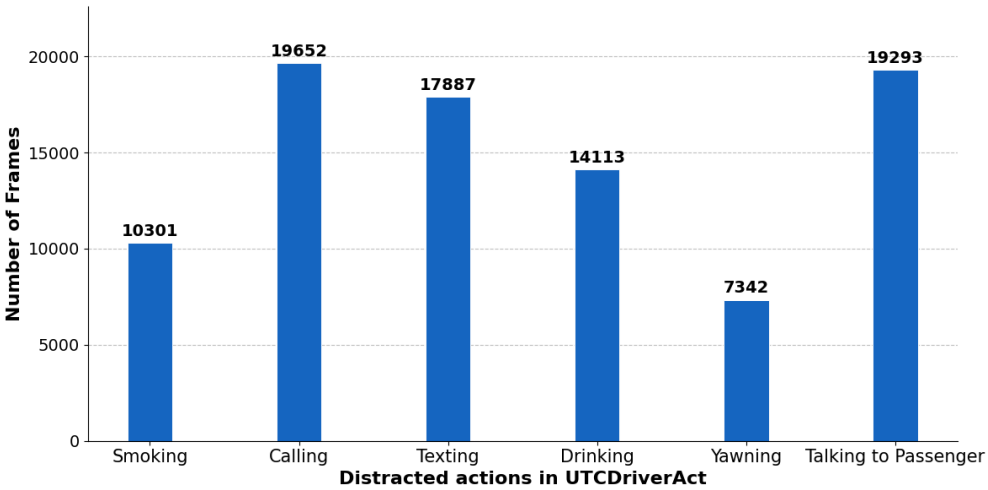


Figure 5. Distribution of annotated frames per action class.

Figure 5 presents the duration of each driver action in UTCDriverAct dataset. From this Figure, we can see that two actions, *Calling* and *Talking to Passenger*, have the highest number of samples, with 19,652 and 19,293 frames respectively. Meanwhile, *Yawning* is the least represented action with a total of only 7,342 frames. This distribution reflects real-world conditions, where *Calling* or *Talking to Passenger* are prevalent activities that lead to significant driver distraction. Additionally, UTCDriverAct dataset also presents specific challenges, such as the class imbalance between activity categories and the presence of relatively small or occluded objects, which create complex classification scenarios for the model.

4.1.2. Measurement metrics

In this work, to assess the effectiveness of the proposed framework for DAR task, two crucial measurement metrics in video recognition and temporal-spatial localization that are Spatio-Temporal Tube IoU (STT-IoU) [15] and Video-mAP are used. The details of these measurement metrics are described below.

Spatio-Temporal Tube IoU (STT-IoU) quantifies the spatio-temporal overlap between a predicted action tube and its corresponding ground-truth tube by aggregating spatial IoU across the temporal dimension. It is considered as the extension of the conventional IoU metric from static images to video sequences and serves as an effective evaluation metric for video understanding tasks. The formulation of STT-IoU is presented in the following equation:

$$STT\text{-}IoU = \left(\frac{Temporal\ \ Intersection}{Temporal\ \ Union} \right) \times Average\ IoU_{2D} \tag{3}$$

Video mean Average Precision (Video-mAP) is also a commonly used evaluation metric for spatio-temporal action recognition in videos. It assesses not only classification performance but also spatial and temporal localization accuracy of predicted actions. Unlike image-based mAP metric, which deals with individual bounding boxes, Video-mAP evaluates action tubes which are the sequences of bounding boxes associated with an action across consecutive frames. Depending on threshold of the STT-IoU metric, a detected action is considered as a true positive or a false negative prediction. From this, a curve is generated to present the relationship between precision and recall metrics at a pre-defined STT-IoU, namely Precision-Recall curve. Average

Precision (*AP*) is computed as the area under Precision-Recall curve and Video-mAP is the average value of *AP* over all examined classes.

$$Video\text{-}mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c, \quad (4)$$

where AP_c is the average precision estimated for class *c* and *C* is the number of action classes.

4.2. Experimental results

In this study, we conduct two experimental scenarios to provide a comprehensive evaluation on the effectiveness of the proposed framework in DAR task. For the first experiment, the proposed framework is applied on UTCDriverAct dataset to present the recognition accuracy for each individual action class and the overall accuracy across all classes. For the second experimental scenario, an ablation study is performed to assess the role of the backbone network for pose estimation model within the proposed framework, considering both recognition accuracy and computational cost. These results also enable recommendations for the deployment of the proposed framework in practical applications. In this work, all experiments are conducted on an Ubuntu 20.04.6 LTS system powered by an NVIDIA GeForce RTX 3060 GPU (12GB VRAM).

4.2.1. Recognition results obtained for UTCDriverAct

Table 1 presents the obtained results of the proposed method HIT4DAR on UTCDriverAct dataset at the STT-IoU threshold of 0.5. The best results are bold, and the second-best ones are underlined. As seen in this Table, we can realize that recognition performance achieves the highest values on *Texting* and *Calling* classes. F1-score@0.5 and AP@0.5 for *Texting* class reach 87.5% and 78.9%, respectively, whereas the corresponding values for *Calling* class are 71.0% and 56.0%, respectively. The worst performance is shown in *Yawning* class, in which F1-score@0.5 and AP@0.5 achieve only 40.9% and 26.8%, respectively. These results can be attributed to the following factors. Firstly, by leveraging both appearance and skeleton data as complementary to each other, HIT model can capture human-object interaction information. Consequently, this model can obtain high performance on several actions which demonstrate explicit human-object interactions, for example *Calling*, *Texting*, *Drinking*. This also explains the limitations in recognition performance observed for the *Yawning* and *Smoking* classes, where fine-grained facial movements and small-scale objects, such as cigarettes, are difficult to capture.

Table 1. Recognition performance of the proposed method HIT4DAR on UTCDriverAct at STT-IoU threshold of 0.5. The best values are shown in bold, and the second-best values are underlined.

Activity	TP	FP	FN	Precision (%)	Recall (%)	F1-score@0.5 (%)	AP@0.5 (%)
Smoking	22	26	24	45.8	47.8	46.8	33.6
Calling	11	4	5	<u>73.3</u>	<u>68.8</u>	<u>71.0</u>	<u>56.0</u>
Texting	14	2	2	**87.5**	**87.5**	**87.5**	**78.9**
Drinking	13	9	9	59.1	59.1	59.1	42.6
Yawning	19	26	29	42.2	39.6	40.9	26.8
Talking to Passenger	15	10	9	60.0	62.5	61.2	48.9

4.2.2. An ablation study on different pose estimation models for DAR

Table 2. Quantitative evaluation of DAR using different pose estimation backbones (STT-IoU = 0.5). The best values are shown in bold, and the second-best values are underlined.

Activity	Keypoint-based model	TP	FP	FN	Precision (%)	Recall (%)	F1-score@0.5 (%)	AP@0.5
Smoking	ResNet-50	22	26	24	45.8	47.8	**46.8**	**33.6**
	ResNet-101	21	23	25	47.7	45.7	46.7	32.4
	ResNet-X-101-32x8d	18	23	28	43.9	39.1	41.4	30.8
	YOLOv8-pose	18	21	28	46.2	39.1	42.4	33.2
Calling	ResNet-50	11	4	5	73.3	68.8	**71.0**	**56.0**
	ResNet-101	11	6	5	64.7	68.8	66.7	54.2
	ResNet-X-101-32x8d	11	4	5	73.3	68.8	**71.0**	53.9
	YOLOv8-pose	11	4	5	73.3	68.8	**71.0**	55.5
Texting	ResNet-50	14	2	2	87.5	87.5	**87.5**	78.9
	ResNet-101	12	2	4	85.7	75.0	80.0	77.4
	ResNet-X-101-32x8d	11	2	5	84.6	68.8	75.9	**84.4**
	YOLOv8-pose	11	2	5	84.6	68.8	75.9	78.8
Drinking	ResNet-50	13	9	9	59.1	59.1	**59.1**	42.6
	ResNet-101	13	11	9	54.2	59.1	56.5	35.4
	ResNet-X-101-32x8d	13	9	9	59.1	59.1	**59.1**	**42.6**
	YOLOv8-pose	10	5	12	66.7	45.5	54.1	39.9
Yawning	ResNet-50	19	26	29	42.2	39.6	40.9	26.8
	ResNet-101	19	24	29	44.2	39.6	**41.8**	27.6
	ResNet-X-101-32x8d	19	25	29	43.2	39.6	41.3	**29.4**
	YOLOv8-pose	16	19	32	45.7	33.3	38.6	24.6
Talking to Passenger	ResNet-50	15	10	9	60.0	62.5	**61.2**	**48.9**
	ResNet-101	15	10	9	60.0	62.5	**61.2**	47.1
	ResNet-X-101-32x8d	14	10	10	58.3	58.3	58.3	44.2
	YOLOv8-pose	14	12	10	53.9	58.3	56.0	41.2

From the pipeline of the proposed framework, we can realize that the quality of the pose estimation model significantly affects the overall recognition performance. Particularly, it is challenging to accurately extract driver keypoints within the vehicle cabin environment due to severe occlusions. In this section, we present an ablation study to investigate the impact of different backbone networks in the pose estimation model on recognition performance in DAR. For this purpose, three deep-learning networks that are ResNet-101, ResNet-X-101-32x8d and YOLOv8-Pose [26] are employed as alternatives to ResNet-50 backbone in the proposed framework for human pose estimation task. While ResNet-101 and ResNet-X-101-32x8d are the invariants of ResNet model, YOLOv8-Pose is the extension of the standard YOLOv8 to simultaneously perform person detection and keypoint localization.

In this experimental scenario, STT-IoU metric is chosen with a pre-defined threshold of 0.5 for all evaluations when working with different backbone networks. This value is consistent with the evaluation protocol established in the ResNet-50 baseline experiments. The obtained

results are summarized in Table 2 using different evaluation metrics, including Precision, Recall, F1-score@0.5, and AP@0.5. By observing this Table, ResNet-50 achieves the highest performance among four examined backbone networks for pose estimation task. ResNet-50 attains the best results on five out of the six examined actions in term of F1-score@0.5 and on three actions according to AP@0.5 metric. On *Texting* activity, it can reach 87.5% and 78.01% in terms of F1-score@0.5 and AP@0.5, respectively. Similar to the obtained results in the first scenario, *Yawning* is the most difficult to recognize, the best performance on this activity is 41.9% for F1-score@0.5 and 20.4% for mAP@0.5. Additionally, in comparison with other pose estimation models, YOLOv8-pose yields the lowest Recall scores across all six driver actions. This can be explained that YOLOv8-pose is sensitive to fine-grained hand movements and strong occlusion in vehicle-cabin, leading significant challenges in driver action recognition task. When using YOLOv8-pose, Precision and Recall metrics reach their highest values at 84.6% and 68.8%, respectively, on *Texting* activity. These values are 2.9% and 18.7% lower than those obtained by ResNet-50 model.

Table 3. Video-AP and mAP results when applying ResNet-50 for pose estimation task. The best values are in bold and the second best are underlined.

Class	ResNet-50	ResNet-101	ResNet-X-101-32x8d	YOLOv8-pose
Smoking	**33.6**	32.4	30.8	<u>33.2</u>
Calling	**56.0**	54.2	53.9	<u>55.5</u>
Texting	<u>78.9</u>	77.4	**84.4**	78.8
Drinking	<u>42.6</u>	35.4	**42.6**	39.9
Yawning	26.8	<u>27.6</u>	**29.4**	24.6
Talking to Pas.	**48.9**	<u>47.1</u>	44.2	41.2
Video-mAP	**47.8**	45.7	<u>47.6</u>	45.5

Table 3 presents the obtained results in terms of Video-mAP and AP@0.5 when using four examined pose estimation models. It is worth noting that Video-mAP is computed across all six driver actions. As shown in this Table, ResNet-50 can reach the highest value of 47.8%, meanwhile YOLOv8-pose exhibits the worst performance at 45.5% on Video-mAP metric. Figure 6 shows an example for key points extraction when applying ResNet-50 and YOLOv8-pose models. From this Figure, ResNet demonstrates more robust and stable performance compared to YOLOv8-pose. ResNet model accurately captures the spatial structure of the main driver's upper body and precisely localizes key joints such as the elbows, wrists, and shoulders during continuous movements (e.g., the action of raising and lowering a cigarette). This more stable pose estimation, accompanied by higher confidence scores, provides higher-quality input features, enabling the subsequent attention module to learn driver–environment interactions more effectively.

Table 4. Performance comparison of Keypoint R-CNN with varying backbones vs. YOLOv8-pose architecture.

Models	**Backbone**	**Inference time (ms)**	**GPU memory (MB)**	**fps**
Keypoint R-CNN	ResNet-50	90.5	2538	11.1
	ResNet-101	119.4	2692	8.4
YOLOv8-pose	ResNeXt-101-32x8d	190.8	2910	5.3

	CSPDarknet (Medium)	14.2	1964	70.5

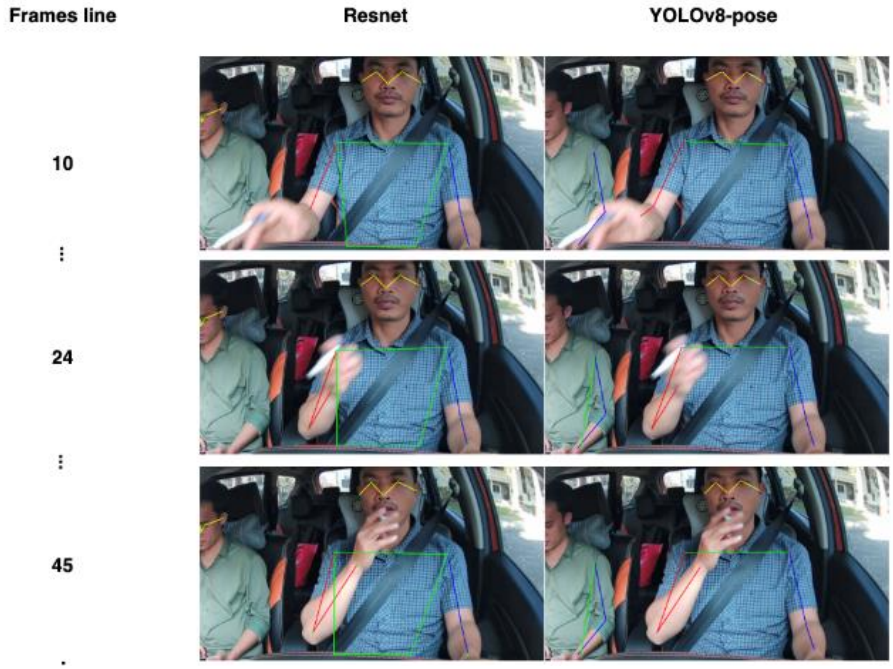


Figure 6. An example for keypoints extraction when using a) ResNet-50 as backbone network of R-CNN model and b) YOLOv8-pose on an image in UTCDriverAct dataset.

Furthermore, we perform an additional evaluation to investigate the trade-off between recognition accuracy and computation time as well as memory usage when using different pose estimation models, shown in Table 4. From an efficiency perspective, YOLOv8-pose outperforms the ResNet variants, achieving a lower inference time of 14.2 ms, a higher processing speed of 70.5 fps, and significantly reduced memory usage of approximately 1964 MB. These values provide a recommendation for research community when deploying the proposed method on a realistic DAR system.

5. CONCLUSIONS AND FUTURE WORKS

In this study, we proposed a comprehensive framework for driver action recognition by leveraging the bi-modal Holistic Interaction Transformer (HIT) network, named HIT4DAR. The main advantage of HIT4DAR is its ability to exploit both visual and skeleton data for driver action representation. Accordingly, the proposed framework aims to learn fine-grained interactions between person, objects, and hands, thereby improving recognition accuracy and robustness to occlusion in the vehicle cabin context. Some extensive experiments are conducted on UTCDriverAct dataset to illustrate the effectiveness of HIT4DAR. Additionally, an ablation study is performed to investigate the influence of pose estimation model on the overall DAR performance. For this target, four models including three variants of ResNet and YOLOv8-pose are used for skeleton extraction step. Experimental results show that the integration of ResNet-50 into HIT4DAR framework leads to superior performance, achieving the highest Video-mAP value of 47.8% over all six evaluated actions. Among the six actions of interest, *Smokin*g and *Yawning* are evaluated as the most challenging to recognize. The AP@0.5 metric on Smoking

and Yawning activities are 33.6% and 26.8%, respectively. These values are much smaller than that on *Texting* activity (78.9%). This gap is the motivation for us to explore another method to improve the performance on all the examined activities. Additionally, we aim to extend the current framework from isolated recognition to continuous recognition, while optimizing the model architecture to ensure effective and smooth real-time deployment on edge devices within intelligent transportation systems. Furthermore, it should be noted that the proposed framework evaluated solely on the UTCDriverAct dataset. Although this is a limitation of our study, it also highlights the need to evaluate the proposed framework on commonly used datasets in future work and compare it with existing approaches.

ACKNOWLEDGMENT

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

REFERENCES

[1]. C. Meurie, O. Lézoray, A comprehensive review of on-board action recognition models in public transportation systems, Expert Systems with Applications 290 (2025) 128311. https://doi.org/10.1016/j.eswa.2025.128311

[2]. Y. Xu, S. Jiang, Z. Cui, F. Su, Multi-view action recognition for distracted driver behavior localization, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2023) 5375-5380. https://doi.org/10.1109/CVPRW59228.2023.00567

[3]. T. Mewborne, L. Zhang, S. Tan, A wearable-based distracted driving detection leveraging BLE, in Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems, (2021) 365-366. https://doi.org/10.1145/3485730.3492872

[4]. 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) 74-77. https://doi.org/10.1145/3507971.3507985

[5]. Hong-Quan Nguyen, Thuy-Binh Nguyen, Trung Kien Tran, Van-Nam Hoang, Thi-Lan Le, Thanh-Hai Tran, Hai Vu, End-to-end deep learning-based framework for driver action recognition, in Proceedings of the IEEE Conference on Multimedia Analysis and Pattern Recognition, (2022) 1-6. https://doi.org/10.1109/MAPR56351.2022.9924944

[6]. Y. Hu, M. Lu, C. Xie, X. Lu, Video-based driver action recognition via hybrid spatial–temporal deep learning framework, Multimedia systems, 27 (2021) 483-501. https://doi.org/10.1007/s00530-020-00724-y

[7]. M. Lu, Y. Hu, X. Lu, Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals, Applied Intelligence, 50 (2020) 1100-1111. https://doi.org/10.1007/s10489-019-01603-4

[8]. H. I. Qamar, U. Saeed, M. Hussain, Driver distraction detection using a multi-stream deep fusion network, in Proceedings of the International Conference on Computing & Emerging Technologies, (2025) 97-104. https://doi.org/10.1007/978-3-031-77617-5_9

[9]. G. Shan, Q. Ji, Y. Xie, Multi-view vision transformer for driver action recognition, in Proceedings of the International Conference on Intelligent Transportation Engineering, (2022) 970-981. https://doi.org/10.1007/978-981-19-2259-6_85

[10]. 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). https://doi.org/10.48550/arXiv.2407.13750

[11]. 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). https://doi.org/10.48550/arXiv.2312.14577

[12]. 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) 5287-5294. https://doi.org/10.1109/CVPRW59228.2023.00557

[13]. G. J. Faure, M.-H. Chen, S.-H. Lai, Holistic interaction transformer network for action detection, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, (2023) 3340-3350. https://doi.org/10.1109/WACV56688.2023.00334

[14]. K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in Proceedings of the 28th International Conference on Neural Information Processing Systems, 1 (2014) 568-576. https://dl.acm.org/doi/10.5555/2968826.2968890

[15]. C. Feichtenhofer, A. Pinz, R. P. Wildes, Spatio-temporal residual networks for video action recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2017) 3468-3476. https://doi.org/10.48550/arXiv.1611.02155

[16]. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016) 770-778. https://doi.org/10.1109/CVPR.2016.90

[17]. 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) 16020-16030. https://doi.org/10.1109/CVPR46437.2021.01576

[18]. 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) 3168-3173. https://doi.org/10.1109/CVPRW56347.2022.00357

[19]. C. Feichtenhofer, X3D: Expanding architectures for efficient video recognition, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2020) 203-213. https://doi.org/10.1109/CVPR42600.2020.00028

[20]. Z. Tong, Y. Song, J. Wang, L. Wang, VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training, in Proceedings of the 36th International Conference on Neural information processing system, 35 (2022) 10078-10093. https://dl.acm.org/doi/10.5555/3600270.3601002

[21]. 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, arXiv preprint arXiv:2010.11929, (2020). https://doi.org/10.48550/arXiv.2010.11929

[22]. Z. Xu, J. Xu, Spatio-temporal decoupling attention transformer for 3D skeleton-based driver action recognition, Complex & Intelligent Systems, 11 (2025) 1-12. https://doi.org/10.1007/s40747-025-01811-1

[23]. C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, (2019) 6202-6211. https://doi.org/10.1109/ICCV.2019.00630

[24]. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 9199.10.5555 (2015): 2969239-2969250. https://doi.org/10.1109/TPAMI.2016.2577031

[25]. K. He, G. Gkioxari, P. Dolla´r, R. Girshick, Mask R-CNN, in Proceedings of the IEEE International Conference on Computer vision, (2017) 2961-2969. https://doi.org/10.1109/ICCV.2017.322

[26]. D. Maji, S. Nagori, M. Mathew, D. Poddar, Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss, in Proceedings of the IEEE/CVF Conference on Computer vision and Pattern recognition, (2022) 2637-2646. https://doi.org/10.1109/CVPRW56347.2022.00297