Quantitative Identification of Driver Distraction: A Weakly Supervised Contrastive Learning Approach

Haohan Yang, Haochen Liu, Zhongxu Hu, Tran Anh-Tu Nguyen, Thierry-Marie Guerra, Chen Lv

Quantitative Driver Distraction Detection: A Supervised Contrastive Learning Approach

Haohan Yang, Haochen Liu, *Graduate Student Member, IEEE*, Zhongxu Hu, *Member, IEEE*, Anh-Tu Nguyen, *Senior Member, IEEE*, Thierry-Marie Guerra, and Chen Lv, *Senior Member, IEEE*

***Abstract*—Accurate recognition of driver distraction is significant for the design of human-machine cooperation driving systems. Existing studies mainly focus on classifying varied distracted driving behaviors, which depend heavily on the scale and quality of datasets and only detect the discrete distraction categories. Therefore, most data-driven approaches have limited capability of recognizing unseen driving activities and cannot provide a reasonable solution for downstream applications. To address these challenges, this paper develops a vision Transformer-enabled supervised contrastive learning framework, in which distracted behaviors are quantified by calculating their distances from the normal driving representation set. The gaussian mixed model (GMM) is employed for the representation clustering, which centralizes the distribution of the normal driving representation set to better identify distracted behaviors. A novel driver behavior dataset and the other three ones are employed for the evaluation, experimental results demonstrate that our proposed approach has more accurate and robust performance than existing methods in recognition of unknown driver activities. Furthermore, the rationality of distraction levels for different driving behaviors is evaluated through driver skeleton poses.**

***Index Terms*—driver distraction quantification, supervised contrastive learning, representation clustering**

I. Introduction

INTELLIGENT driving has attracted considerable attention in recent years, and its development is of great importance to driving safety [1]–[4]. Both naturalistic driving data and in-lab simulator experiments have demonstrated that driver distraction is a leading inducement of traffic a ccidents, and therefore, it is significant to parse driver behaviors for avoiding potential unsafe maneuvers [5]–[7]. For instance, warning signals can be generated to alert distracted drivers to allocate their attention toward possible hazards in advance. Additionally, an adaptive takeover scheme can be designed for various driver states to ensure a smooth and safe control transition [8].

Driver distraction is classified into two categories generally, i.e., physical and cognitive. Accordingly, different modal information, such as vehicle states, electroencephalography (EEG), head/eye movements, etc., is employed to recognize driver distraction. In [9], [10], vehicle states and EEG signals were utilized for inferring cognitive distraction, respectively, but these methods are difficult to achieve satisfactory performance in practice due to channel noises and artifacts. Currently, most studies focus on physical distraction recognition with greater practical significance. Vision-based approaches are widely used in physical distraction detection, for example, estimated head poses and eye gaze directions from raw images were adopted for non-driving activity recognition [11]–[13]. Based on the image and video information, extensive studies also have been carried out on end-to-end driver distraction inference to further reduce the computation cost [14], [15].

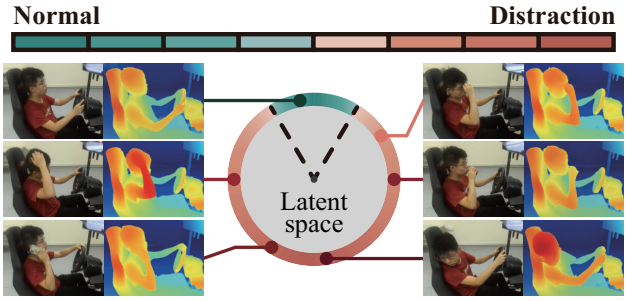

Fig. 1. Driver distraction quantification using the distance from normal behaviors in the latent space. The demonstration can be found on the GitHub website.

In previous studies, driver distraction detection is basically regarded as a classification problem, thus varied supervised learning approaches are developed to tackle it. In [16], a radial-basis neural-network-based framework is established to distinguish distracted behaviors. Using colored depth images, a feed-forward neural network (FFNN) and a support vector machine (SVM) were built to identify driver activities, respectively [17], [18]. By comparing normal driving parameters against distracted ones, a fuzzy logic algorithm is proposed to recognize driver distraction [19]. Also, an attention-based long short-term memory (LSTM) network architecture was utilized for detecting driver distraction through multi-modal driving data [20]. Since convolutional neural networks (CNN) can extract image/video feature representations better and faster compared with the above methods, related approaches have been proposed from different perspectives [21]–[29]. In [23], the pre-trained CNN model was employed to identify driver behaviors using segmented images. To further improve the inference efficiency, several methods are proposed to reduce the model size. For instance, a depthwise separable convolution approach was adopted to establish a lightweight CNN model for driver activity recognition [27]. Furthermore, 3D CNN models have also been designed for extracting the motion information hidden in video frames. In [29], a dual-stream 3D residual network was proposed to enhance spatio-temporal feature representations and improve the non-driving activity recognition performance. However, the aforementioned tech-

niques require massive samples with labels, especially for distracted driving ones, which is laborious and costly. Also, it is prohibitively difficult to contain all types of driver behaviors in manually collected datasets, thus the recognition ability of models is limited for unseen activities previously.

To enable models to better identify unknown distracted behaviors, some semi-supervised and unsupervised learning approaches have been presented. In [30], a Laplacian SVM was employed for driver distraction detection using eye and head movements. Based on the multi-modal information, such as electromyography (EMG), galvanic skin responses, etc., an unsupervised network was designed for recognizing driver distraction [31]. Whereas both above schemes require specific hardware equipment with high costs, a contrastive learning framework was employed to identify driver distraction according to raw images obtained by inexpensive cameras [32]. Driver distraction levels are still discrete in these studies, thus cannot provide a practical solution for downstream applications, such as the shared control/planning scheme design [33], [34], etc. Furthermore, a few contrastive learning methods were designed to quantify driver anomaly, which bring an up-to-date perspective to the research on driving monitoring systems [35], [36]. Nevertheless, reported studies on driver distraction quantification are still quite limited.

Compared with previous studies, the main contributions of this article are summarized into three aspects:

1) A vision Transformer-enabled supervised contrastive learning framework is developed to recognize distracted driving and quantify driver distraction levels, which suggests a viable generic technique for driver monitoring.

2) The Gaussian mixed model (GMM) is employed for representations clustering of normal driving activities, which further enhances the model representation capability of detecting unknown distracted behaviors.

3) A novel driver behavior dataset is constructed to evaluate the proposed method and other state-of-the-art methods. Also, driver skeleton poses are extracted to validate the rationality of obtained distraction levels.

The remainder is organized as follows: Section II describes the structure and the training process of the developed framework. The experimental protocol of our proposed driver behavior dataset and its feature comparison with others are illustrated in Section III. In Section IV, classification results and distraction quantification evaluations are presented and analyzed to demonstrate our model's superior performance. Finally, conclusions and some further works are summarized in Section V.

II. Methodology

In this section, we describe the proposed contrastive representation learning framework. The problem of driver distraction quantification is illustrated firstly, then the architecture of our model and its training procedure are introduced, respectively.

A. Problem Formulation

Driver behaviors can basically be classified into two categories, i.e., normal and distraction. Normal behaviors are generally quite similar, while distracted ones can be varied during driving [15]. A representation set of normal driving is constructed accordingly, and driver distraction levels can be obtained by calculating distances between the given driving activities and the set of normal ones in the latent space. Also, any activity with a distraction level beyond the threshold is detected as distracted driving behavior. The above conception requires a model to align feature representations of normal driving activities and minimize the similarity in representations between distracted behaviors and normal ones. This goal can be formulated as

$$\begin{gathered} \left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2<<\left\|\boldsymbol{v}_i-\boldsymbol{v}_k\right\|_2 \\ i, j \in \mathcal{X}\,(i \neq j),\ k \in \mathcal{D} \end{gathered} \tag{1}$$

wherein $\boldsymbol{v}$ is the feature representation of the corresponding samples, $\mathcal{X}$, $\mathcal{D}$ represent index sets of normal and distraction samples, respectively.

B. Model Construction

The architecture of the developed model is shown in Fig. 2, which consists of five parts, i.e., data augmentation, encoder, decoder, projection and loss design.

• Data augmentation

In practice, cameras' installation positions/angles, signal noises and the ambient light inevitably change under various driving environments. Therefore, four methods, including rotation, cropping, noise enhancement and color jitter, are employed for the data augmentation. During the model training, a combination of these four random augmentations is applied to each input image for generating a corresponding image pair. Also, all input images (W, H, C) are resized to (224, 224, C) in this part. For convenience, the augmented index sets of normal (distracted) samples in a mini-batch and the training sets are denoted as $\mathcal{X}^a$ ($\mathcal{D}^a$) and $\mathcal{X}^t$ ($\mathcal{D}^t$), respectively.

• Encoder

A hierarchical vision Transformer using shifted windows, namely Swin Transformer (Swin-T), is employed as a backbone encoder for extracting feature representations of images [37]. As shown in Fig. 2, a Swin-T block consists of multi-head self-attention modules with regular and shifted windowing configurations, denoted as W-MSA and SW-MSA, respectively, followed by the multilayer perceptron (MLP), and the LayerNorm (LN) layer with a residual connection is applied before each module. The Swin-T model transfers each input image as follows,

$$\begin{cases} \hat{\mathbf{z}}^l=\text{W-MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right)+\mathbf{z}^{l-1} \\ \mathbf{z}^l=\text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^l\right)\right)+\hat{\mathbf{z}}^l \\ \hat{\mathbf{z}}^{l+1}=\text{SW-MSA}\left(\text{LN}\left(\mathbf{z}^l\right)\right)+\mathbf{z}^l \\ \mathbf{z}^{l+1}=\text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right)+\hat{\mathbf{z}}^{l+1} \end{cases} \tag{2}$$

where $\hat{\mathbf{z}}^l$ and $\mathbf{z}^l$ are output features of the MSA-based modules and the MLP module for block l, respectively. The model produces a hierarchical feature map (7, 7, 768) at last, and its corresponding representation $\boldsymbol{h} \in \mathbb{R}^{768}$ is obtained by applying a global average pooling (AP) layer.

• Decoder

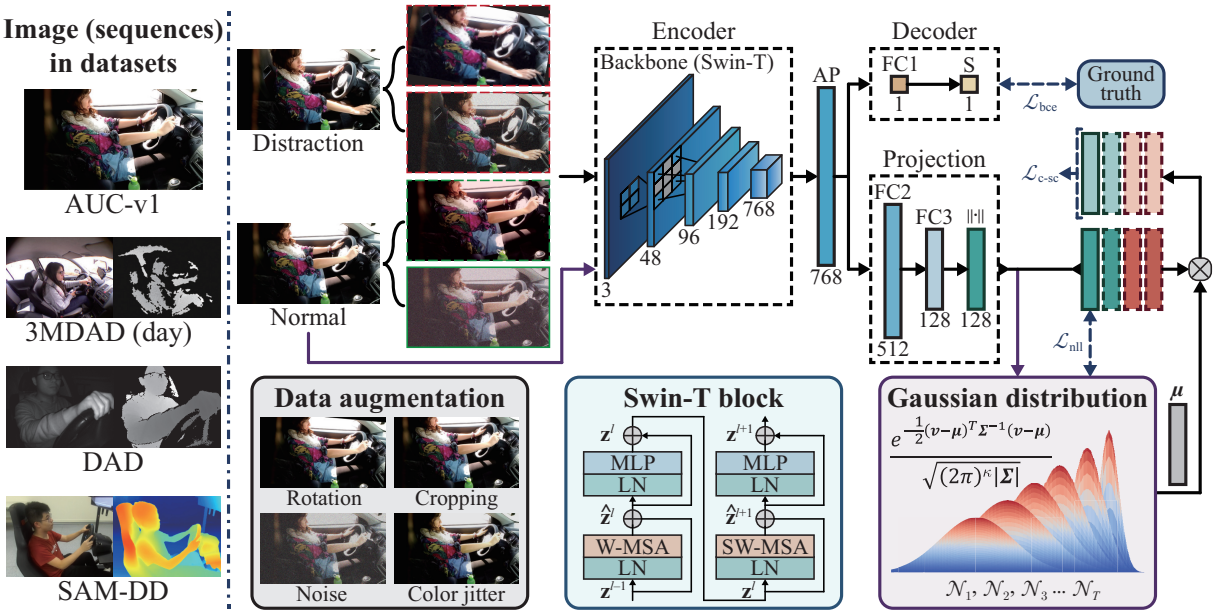

Fig. 2. Overview of the developed contrastive representation learning approach for quantitative driver distraction detection.

A binary classification decoder is designed to further improve the model's feature capturing capability. The fully connected (FC) layer and a Sigmoid (S) activation function are performed to transform the feature representation $\boldsymbol{h}$ into a constant $c \in [0, 1]$, i.e., the probability of the normal driving behavior.

• Projection

A deeper network structure can filter unnecessary image information for contrastive learning tasks [38]. In this study, two fully connected layers and the $\ell 2$ normalization are conducted to transform $\boldsymbol{h}$ into an embedding $\boldsymbol{v} \in \mathbb{R}^{128}$. Accordingly, all input images are mapped on a unit hypersphere through the projection.

• Loss design

Three loss functions, i.e., binary cross-entropy (BCE) loss, clustering-based supervised contrastive (C-SC) loss and negative log-likelihood (NLL) loss, are designed in this study. Cross-entropy loss is generally utilized for the classification of distracted behaviors in previous studies, whereas in our model, the BCE loss is only employed for assisting in better capturing feature representation of normal driving activities. The BCE loss is defined as

$$\mathcal{L}_{\text{BCE}} = \sum_{i \in \mathcal{X}^a \cup \mathcal{D}^a} \left[y_i \cdot \ln c_i + (1 - y_i) \cdot \ln (1 - c_i) \right] \tag{3}$$

wherein y_i and c_i denote the label of the ith sample and its predicted probability of normal driving, respectively. It is noted that the label $y_i \in \{0, 1\}$, in which "0" and "1" represent normal and distracted driving, respectively.

Normal driving behaviors are generally similar in practice, and therefore, the distribution of their feature representations is expected to be concentrated. Based on the conception of GMM clustering, a multivariate Gaussian distribution is constructed using representations of normal driving samples in the training dataset. To enable normal samples to cluster, minimizing both distances from the embedding representations to their center and the covariance of the distribution is desired. Accordingly, the NLL loss is constructed as

$$\begin{aligned} \mathcal{L}_{\text{NLL}} &= -\sum_{i \in \mathcal{X}^a} \ln \left[\mathcal{N} \left(\boldsymbol{v}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) \right] \\ &= -\sum_{i \in \mathcal{X}^a} \ln \left[\frac{1}{\sqrt{(2\pi)^{\kappa} |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2} (\boldsymbol{v}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{v}_i - \boldsymbol{\mu})} \right] \end{aligned} \tag{4}$$

where $\mathcal{N}$ is the multivariate Gaussian distribution of the representation set at each training epoch, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and the covariance of the distribution, respectively, $\kappa = 128$ is the dimension of each representation $\boldsymbol{v}$.

The supervised contrastive (SC) loss is designed to pull together the representations belonging to the same class in the latent space, which is formulated below [39],

$$\mathcal{L}_{\text{SC}} = -\sum_{i \in \mathcal{X}^a \cup \mathcal{D}^a} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \ln \frac{\exp \left(\boldsymbol{v}_i \cdot \boldsymbol{v}_j / \tau \right)}{\sum_{k \in \mathcal{A}(i)} \exp \left(\boldsymbol{v}_i \cdot \boldsymbol{v}_k / \tau \right)} \tag{5}$$

where $\mathcal{P}(i)$ is an index set that has the same label as embedding $\boldsymbol{v}_i$, and $|\mathcal{P}(i)|$ denotes its cardinality, $\mathcal{A}(i) \equiv \mathcal{P}(i) / \{i\}$ is a relative complement of the index, and $\tau \in \mathbb{R}^+$ is a scalar temperature parameter. To further align normal driving activities, this study translates the original hypersphere center to the center of representation distribution $\mathcal{N}$. Consequently, the embedding representation in the translated latent space and the C-SC loss are, respectively, described as

$$\boldsymbol{v}^c = \frac{\boldsymbol{v} - \boldsymbol{\mu}}{\|\boldsymbol{v} - \boldsymbol{\mu}\|_2} \tag{6}$$

$$\mathcal{L}_{\text{C-SC}} = -\sum_{i \in \mathcal{X}^a \cup \mathcal{D}^a} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \ln \frac{\exp \left(\boldsymbol{v}_i^c \cdot \boldsymbol{v}_j^c / \tau \right)}{\sum_{k \in A(i)} \exp \left(\boldsymbol{v}_i^c \cdot \boldsymbol{v}_k^c / \tau \right)} \tag{7}$$

Eventually, the training loss is summarized as

$$\mathcal{L}_{\text{S}} = \alpha_1 \mathcal{L}_{\text{BCE}} + \alpha_2 \mathcal{L}_{\text{NLL}} + \alpha_3 \mathcal{L}_{\text{C-SC}} \tag{8}$$

where α_1, α_2, α_3 are weights of the corresponding loss function.

The pseudo-code of our model's learning strategy is provided in Algorithm 1. A transfer learning technique is applied to fine-tune the pre-trained Swin-T model, which transfers the domain knowledge learned from large-scale datasets to driver distraction detection tasks. The proposed model is trained with an Adam optimizer, and the learning rate is 0.0001. The mini-batch size is selected as 128 with 20 training epochs, and the model with the highest detection accuracy is saved. The temperature parameter τ is set to 0.07, and the loss weights $\alpha_1 = \alpha_3 = 1$, $\alpha_2 = 0.001$. The whole network is developed with PyTorch (https://pytorch.org/).

Algorithm 1: Learning strategy of the driver distraction detection framework

Input: training sets $\mathcal{X}^t$, $\mathcal{D}^t$, temperature parameter τ, loss weights α_1, α_2, α_3
for $t \leftarrow 1$ **to** *max epoch* **do**
  # Stage 1: obtain the distribution $\mathcal{N}$
  Initialize the normal driving representation set $\mathcal{V}$;
  for $i \leftarrow 1$ **to** $|\mathcal{X}^t|$ **do**
    Calculate normal embeddings $\boldsymbol{v}_i$;
    Append $\boldsymbol{v}_i$ into $\mathcal{V}$;
  end
  Calculate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the distribution $\mathcal{N}_t(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;
  # Stage 2: train the proposed model
  for *Gradient_step* $\leftarrow 1$ **to** *max iteration* **do**
    Generate augmented image sets $\mathcal{X}^a$ and $\mathcal{D}^a$;
    Calculate loss according to Eq. (8);
    Update f_{Dec} by $\mathcal{L}_{\mathrm{BCE}}$;
    Update f_{Proj} by $\mathcal{L}_{\mathrm{C-SC}}$, $\mathcal{L}_{\mathrm{NLL}}$;
    Update f_{Enc} by $\mathcal{L}_{\mathrm{S}}$;
  end
end

To ensure the stability of calculated distraction levels, the k-nearest neighbor (KNN) algorithm with 2 neighbors is utilized for distraction quantification,

$$\zeta_i = \min\left(\frac{\|v_i - v_j\|_2 + \|v_i - v_k\|_2}{2}\right) \quad i \in \mathcal{X} \cup \mathcal{D},\ j,k \in \mathcal{X}^t\ (j \neq k) \tag{9}$$

wherein ζ_i is the distraction level of ith sample.

III. Experiments

In this section, we illustrate the experimental scheme in detail and compare the features of four publicly available driver distraction datasets.

A. Experimental Protocol

Few high-quality vision-based driver distraction datasets are publicly available, and most of them are collected inside the vehicle cabin. To guarantee safety and efficiency, most intelligent driving strategies are initially verified through hardware-in-the-loop tests, rather than field experiments. In this study, we have constructed a **S**ingapore **A**uto**M**an@NTU **d**istracted **d**riving (SAM-DD) dataset, which can be migrated to studies involving driver states based on driving simulators, e.g., driving authority allocation, etc.

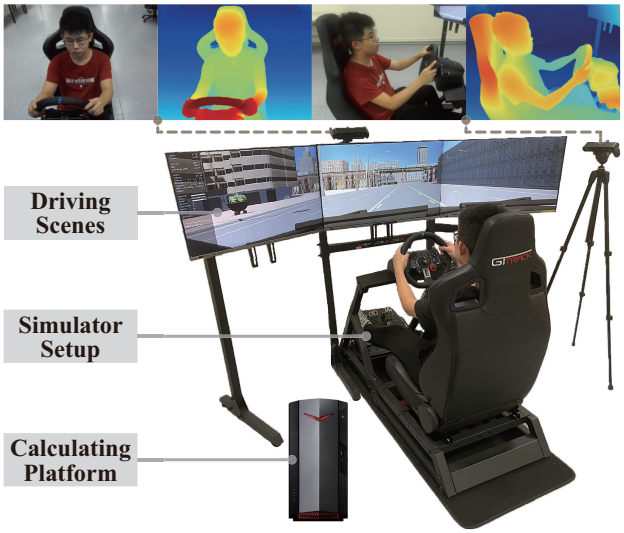

Fig. 3. Driver-in-the-loop experimental platform.

The driver-in-the-loop experimental platform, as shown in Fig. 3, is composed of a physical simulator (Logitech G29), two cameras (Zed 2i) and a computer (NVIDIA GTX 2080 Ti with 32 GB RAM). Experimental data are collected from 42 participants (34 males, 8 females) with varied ages and driving experience. Through selecting and integrating abnormal behaviors in previous studies, nine representative physical distracted behaviors are recorded in our dataset, including eight non-driving activities (i.e., drinking, talking left/right, texting left/right, touching hairs, adjusting glasses, reaching behind) and one fatigue-related behavior (i.e., head dropping). The datasets are collected in synchronized RGB and depth modalities with a resolution of 1200×900 pixels. In addition to the lateral camera used to detect head and arm movements, a front camera is installed to capture drivers' facial information. An example of two modalities and two views is presented in Fig. 3.

B. Datasets Features

Table I illustrates the different characteristics of four publicly available driver distraction datasets, including the first version of the American University in Cairo Distracted Driver (AUC-v1) dataset, a multi-view, multimodal and multispectral Driver Action Dataset (3MDAD), the Driver Anomaly Detection (DAD) dataset and our proposed dataset [18], [22], [35]. AUC-v1 is the first public driver distraction dataset, consisting of 17308 samples from 31 participants. In the experiments, ten typical driver behaviors, including safe driving, are recorded by a roof handle-mounted camera. 3MDAD contains two natural driving sets collected during the daytime and the night, respectively, and only the daytime one is employed in this study. 3MDAD (day) provides temporally synchronized RGB and depth frames with front and side views, in which 16

TABLE I
CHARACTERISTICS COMPARISON ACROSS DRIVER BEHAVIOR DATASETS.

Datasets	AUC-v1 [22]	3MDAD (day) [18]	DAD [35]	SAM-DD [ours]
Resolution ($W \times H$)	1920×1080	640×480	224×171	1200×900
Image modes	RGB	RGB & Depth	Infrared & Depth	RGB & Depth
Sample types	Single frame	Single frame	Video clip	Single frame
Labeled behaviors	10	16	9	10
Sample sizes	17308	111017	67051	51175
Participants	31	50	31	42
Gender ratio (M / F)	22 / 9	38 / 12	20 / 11	34 / 8
Collection scenes	Parked vehicle	Natural driving	Parked vehicle	In-lab
Views	1	2	2	2

driving actions from 50 subjects are recorded. However, the depth images in the 3MDAD are difficult to be utilized for driver behavior detection directly due to their low quality. It is noted that the 3MDAD is an imbalanced dataset wherein normal driving samples only account for 1/16. DAD dataset has synchronized infrared and depth video frames from both front and top views, and each input clips consist of 16 frames. Especially, 16 unlabeled abnormal driving activities in the test set are unavailable in the training one, which requires the model can recognize previously unseen distracted behaviors.

Features of our dataset are summarized below,

• The SAM-DD dataset is large enough for training learning-based models from scratch. Also, researchers can conveniently migrate the trained model to targeted downstream tasks.

• The SAM-DD dataset contains high-quality multi-modal information, i.e., RGB and depth, which can improve the model's reliability against various driving environments. The dataset has multiple views, which are recorded synchronously and complement each other. Accordingly, researchers can utilize the dataset for wider driver states-related tasks. (Note: Only the lateral view is employed in this study)

• The SAM-DD dataset is mainly for intelligent driving research in the laboratory, including driving takeover systems, remote driving, and control strategies involving driver states, etc.

The selected four datasets contain many different types of distracted driving activities, which can be conveniently utilized to test models' capability of recognizing previously unseen distracted behaviors. Also, these datasets are complementary in terms of image modes, sample types, data distributions and collection scenes. Different characteristics of the datasets can enable the developed driver distraction quantification model to be comprehensively evaluated.

IV. RESULTS AND DISCUSSIONS

In this section, we first evaluate the proposed model and the other state-of-the-art methods with varied backbones and loss functions. Then, the clustering characteristics of varied approaches are compared. Finally, we analyze the distributions of driver distraction levels obtained by the designed model and evaluate its rationality using drivers' skeleton key points.

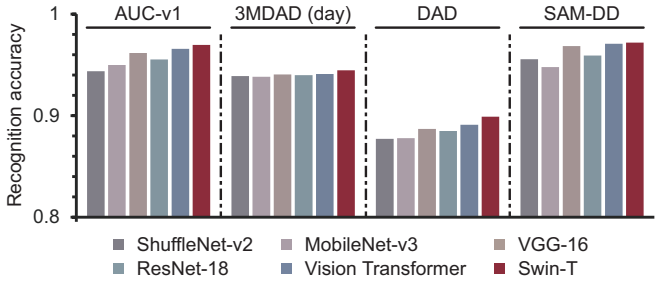

Fig. 4. Comparison of varied backbones over recognition accuracy across different driver behavior datasets.

A. Classification performance

To verify the feature extraction capability of the Swin-T in the proposed framework, varied state-of-the-art models are employed as encoders for comparison, and the recognition results across different datasets are shown in Fig. 4. Although the recognition accuracy of each backbone varies across the datasets with different data distributions and characteristics, the Swin-T obviously outperforms other backbones in all cases, indicating it can capture the feature representation of driving behaviors more effectively. Especially, a 3D Swin-T model is selected to extract representations of sequential frames in the DAD dataset. Since this study aims to establish an efficient contrastive learning framework, but not elaborately design a classification model, more detailed discussions of varied encoders' recognition results are omitted here.

To better demonstrate the superiority of our framework, two other baseline methods are designed, i.e.,

• E2E-Sup: an end-to-end supervised learning approach. The method is utilized to classify driver behaviors into normal and distraction through their probabilities calculated by a binary classifier.

• SupCon: a supervised contrastive learning approach. The method can recognize driving behaviors by quantifying driver distraction levels. One significant difference from our proposed approach is that it lacks clustering loss $\mathcal{L}_{\mathrm{NLL}}$ during training.

To achieve a fair comparison, the Swin-T is employed as the encoder in all approaches. All models are trained by a varied number of distracted activities to investigate their capability of recognizing previously unseen activities. Experimental results are analyzed in terms of three evaluation metrics, i.e., accuracy, area under curve (AUC), and F1-score, as described in Table II.

TABLE II
RECOGNITION PERFORMANCE OF MODELS TRAINED BY THE VARIED NUMBER OF DISTRACTION ACTIVITIES ACROSS DIFFERENT DRIVER BEHAVIOR DATASETS.

Distractions number	E2E-Sup			SupCon			SupCon+GMM (ours)		
	Accuracy	AUC	F1-scores	Accuracy	AUC	F1-scores	Accuracy	AUC	F1-scores
AUC-v1									
2	0.7184	0.8638	0.7865	0.9399	0.9577	0.9625	**0.9408**	0.9622	0.9627
4	0.8269	0.9081	0.8795	0.9353	0.9525	0.9595	**0.9362**	0.9680	0.9596
6	0.8882	0.9497	0.9256	**0.9424**	0.9687	0.9635	0.9418	0.9618	0.9634
9	**0.9709**	0.9908	0.9815	0.9631	0.9771	0.9799	0.9697	0.9778	0.9808
3MDAD (day)									
3	0.9024	0.7777	0.9480	0.9397	0.6366	0.9688	**0.9399**	0.7818	0.9687
7	0.9090	0.7424	0.9515	**0.9392**	0.6916	0.9686	0.9390	0.5730	0.9685
11	0.9319	0.7889	0.9583	0.9403	0.6161	0.9691	**0.9417**	0.7063	0.9697
15	0.9403	0.7427	0.9691	0.9415	0.6299	0.9690	**0.9447**	0.7363	0.9712
DAD									
2	0.7908	0.8277	0.5644	0.8552	0.8989	0.7693	**0.8584**	0.9181	0.7983
4	0.8036	0.8291	0.6077	0.8637	0.8983	0.7948	**0.8665**	0.9214	0.7954
6	0.8317	0.8791	0.6784	0.8717	0.9133	0.8066	**0.8730**	0.9232	0.8208
8	0.8517	0.8921	0.7379	0.8971	0.9342	0.8382	**0.8989**	0.9321	0.8453
SAM-DD									
2	0.8639	0.9536	0.8618	0.9274	0.9668	0.9334	**0.9300**	0.9647	0.9353
4	0.8945	0.9588	0.8991	0.9313	0.9705	0.9374	**0.9326**	0.9682	0.9380
6	0.9355	0.9795	0.9406	0.9432	0.9778	0.9486	**0.9451**	0.9774	0.9510
9	**0.9733**	0.9914	0.9757	0.9641	0.9828	0.9678	0.9720	0.9781	0.9746

It is seen that the recognition performance of varied approaches is close across AUC-v1, 3MDAD (day), and SAM-DD datasets when the models are trained with all distraction samples, and the E2E-Sup can slightly outperform the SupCon-based methods in some datasets at this point. However, the E2E-Sup model's performance decreases dramatically with the decline of the varied number of distraction behaviors during training, and its reduction varies with the different sample distributions in the datasets. For instance, the maximum reduction of the E2E-Sup model's accuracy is nearly 26.0% in the AUC-v1 dataset, while only about 4.0% in the 3MDAD (day) dataset due to the imbalanced sample distribution. In comparison, the reductions in the accuracy of our developed model are only almost 3.0% and 0.6% in AUC-v1 and 3MDAD (day) datasets, respectively. Since the testing set of the DAD dataset has 16 additional distracted activities that do not appear in its training set, the performance of the E2E-Sup trained by all training samples is always worse than the SupCon-based models. Obviously, accuracy reductions of the SupCon-based methods are significantly lower than that of the E2E-Sup method when the training is conducted with partial distracted driving activities, which demonstrates that SupCon-based methods are greatly robust in identifying unknown driving behaviors. Interestingly, SupCon-based models trained with more distracted behaviors show relatively poorer recognition results in some cases, on account of the more complex segmentation boundary generated by more driving behaviors is hard to be fully learned in practice. Furthermore, the GMM-based representation clustering is conducted in our developed approach, which improves the model performance in most cases and requires no additional computations compared with the SupCon. Clustering-based SupCon improves the AUC by 12.8%–18.6% compared with the SupCon in 3MDAD (day) dataset, indicating that the GMM-based representation clustering can significantly enhance the model's robustness in the decision threshold selection, especially in the case of few training samples of normal driving.

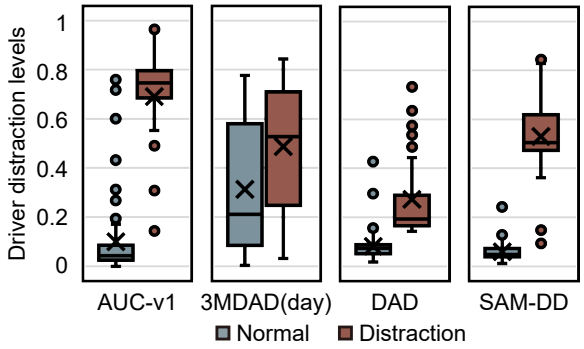

Fig. 5. Distributions of driver distraction levels calculated by the proposed approach on different test sets.

B. Distraction quantification evaluations

In addition to efficiently recognizing previously unknown driving behaviors, our proposed approach aims to properly reflect distraction levels of different driver activities. Distributions of distraction levels calculated by the developed approach on the test sets of four datasets employed in this study are shown in Fig. 5. It is seen that the datasets containing RGB images, i.e., AUC-v1, 3MDAD (day), and SAM-DD, have a greater distribution of the data domain than the DAD dataset. Due to numerous mislabeled samples, normal behaviors and distracted ones have a significant overlap in the 3MDAD (day)

dataset, indicating that precisely labeled data are required by the proposed approach for constructing the distribution of normal driving actions during training. Our approach distinctly split normal and distraction samples on the AUC-v1, DAD, and SAM-DD datasets except for a few outliers. Specifically, the distraction levels of normal samples are low and concentrated, while the distraction levels of distracted behaviors are dispersedly distributed and separated from the normal ones. Furthermore, an obvious distinction between the normal and distraction samples, especially in the AUC-v1 and SAM-DD datasets, demonstrates our method provides a loose range of the decision threshold for recognizing driving behaviors.

Driver skeleton poses are employed to further assess the rationality of the distraction levels for different distracted driving behaviors. A state-of-the-art human body pose detector is utilized to extract skeleton keypoints, and all keypoints are reshaped and normalized into a vector to compute the distances of different distracted driving behaviors from the normal one [40]. The average distraction levels of varied distracted actions calculated by the different methods are presented in Fig. 6, wherein the distraction levels reflected by the skeleton keypoints conform to human intuition and thus are employed as the reference. It is noteworthy that an approach that can reasonably describe the differences in varied distracted behaviors is regarded as an ideal one, and whether the calculated distraction levels equal the reference is meaningless. From the perspective of graphics, the approach that is more similar to the reference model in terms of the polygon shape can better reflect the distraction levels of driving behaviors.

It is observed that the difference in the distraction levels of varied driving behaviors, excluding "head dropping", calculated by the E2E-Sup method is little in the AUC-v1, 3MDAD (day), and SAM-DD datasets, whereas a relatively obvious distinction is achieved in the DAD dataset. The "head dropping" is an exception since the E2E-Sup wrongly classifies it as a normal behavior, especially in the SAM-DD dataset. The phenomenon indicates that the E2E-Sup method cannot distinguish different distracted behaviors based on a single frame, while the spatio-temporal information can improve its performance in this regard. For the reference model, the distraction level of "texting left" is larger than both "texting right" and "toughing hairs & makeup" in the AUC-v1 dataset, the same phenomenon is presented in our proposed model while an opposite one appears in the SupCon method. Similar examples can also be found in other datasets, which demonstrates that the distraction levels obtained by the clustering-based SupCon are most consistent with the skeleton pose results and our developed model can better reflect the distraction levels compared with other baseline approaches. Interestingly, the distraction levels of symmetrical driving activities, such as "phoning right" and "phoning left", are different since the side-mounted camera makes some driving actions to be more obvious than their symmetrical ones in the captured images.

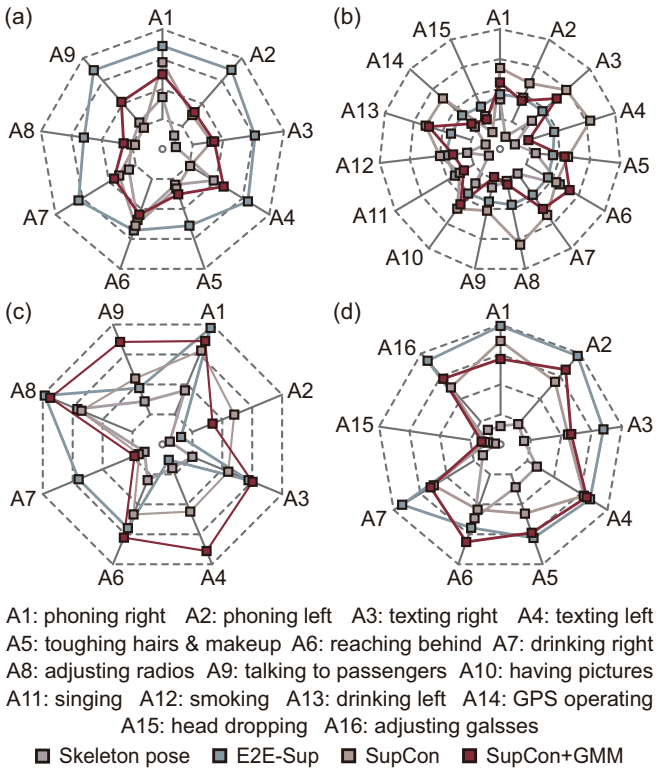

Fig. 6. Comparison of the average distraction levels of different distracted driving behaviors calculated by different approaches in the four test sets. (a) AUC-v1. (b) 3MDAD (day). (c) DAD. (d) SAM-DD.

C. Clustering analysis

To intuitively understand the difference between the baseline methods and ours, a t-distributed stochastic neighbor embedding (t-SNE) approach is employed to visualize the feature representations mapped by different methods [41]. The visualization results across the four datasets are presented in Fig. 7. It is obvious that the representations of normal samples and distracted ones extracted by the E2E-Sup are entangled in all datasets. Conversely, the SupCon-based approaches present distinguishable clustering results in the AUC-v1, DAD, and SAM-DD datasets. Also, representations of normal driving samples generated by SupCon-based approaches in the AUC-v1 appear to be more concentrated than those in the DAD and SAM-DD datasets due to the varied proportions of normal and distraction samples in the test sets. For instance, the normal samples in the test set of AUC-v1 only account for nearly 1/4, while the size of normal samples is almost twice that of distracted driving in the DAD dataset. For the 3MDAD (day) dataset, although the normal samples' representations extracted by the clustering-based SupCon are more distributed on the right side, all methods cannot yield recognizable clusters since many distracted behaviors are mislabeled as normal driving. Furthermore, the clustering results of our developed model provide clearer distinctions between the normal and distracted driving samples compared with the SupCon method, which is beneficial for quantifying driver distraction levels and enhancing the model's recognition performance.

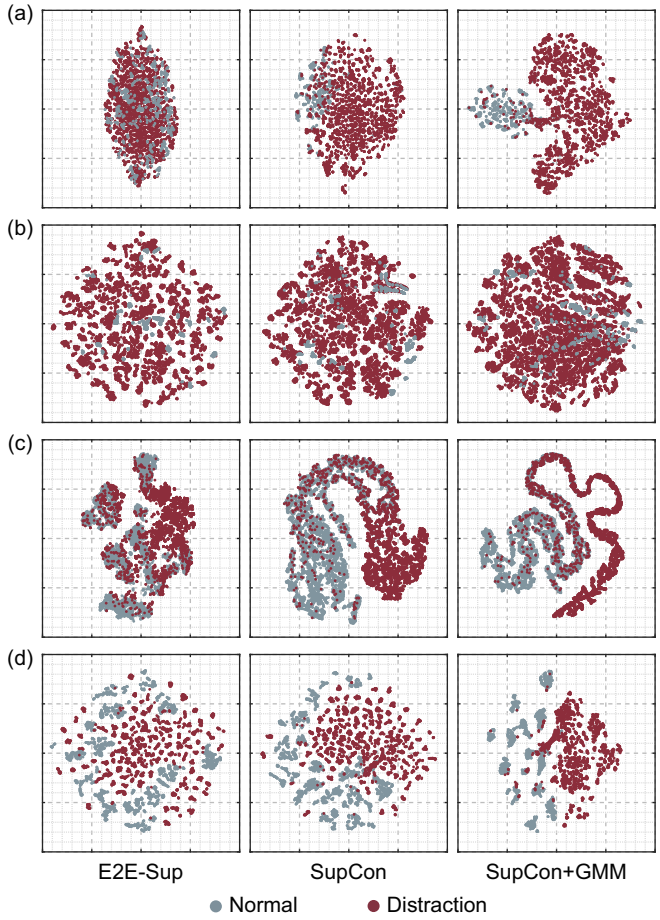

Fig. 7. Visualization results of the feature representations extracted by different approaches across four datasets. (a) AUC-v1. (b) 3MDAD (day). (c) DAD. (d) SAM-DD.

V. Conclusions

Our study aims to tackle the two limitations of most learning-based driving distraction detection models in practical application, i.e.,

- Existing datasets cannot cover all natural driving behaviors, which leads to a drastic deterioration of most models' performance in recognizing unknown driving behaviors.
- Most previous studies regard driving distraction detection as a multi-classification task, but only the recognition of discrete driving behavior categories is not suitable for the development of downstream applications.

In this paper, we propose a driver distraction quantification framework for detecting distracted activities and construct a novel driver behavior dataset for in-lab driving research. A vision transformer-enabled supervised contrastive learning model is designed to recognize distracted driving activities, especially previously unseen ones. Also, a GMM is employed to build the clustering-based representation set of normal driving that is utilized to calculate driver distraction levels. Experimental results across four datasets demonstrate that the designed contrastive learning approach is strongly robust in recognizing unknown distracted activities compared with the E2E-Sup model, and the representation clustering technique enables the model to better reflect distraction levels related to the variation of driver skeleton information. The superior performance of our developed distraction detection framework demonstrates that it is a more practical solution for downstream applications.

In further research, a series of driver states-related intelligent driving schemes, such as takeover systems and shared control strategies, etc., can be optimized and developed through combining our developed distraction detection framework. Also, an important further study is to better align the distraction levels of driving behaviors with their potential risks.

References

[1] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, Apr. 2020.

[2] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep Learning-Based Vehicle Behavior Prediction for Autonomous Driving Applications: A Review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 33–47, Jan. 2022.

[3] C. Huang et al., "Human-Machine Cooperative Trajectory Planning and Tracking for Safe Automated Driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12050–12063, Aug. 2022.

[4] J. Wu, Z. Huang, W. Huan, and C. Lv, "Prioritized Experience-Based Reinforcement Learning with Human Guidance for Autonomous Driving," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 10, 2022, doi: 10.1109/TNNLS.2022.3177685.

[5] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 596–614, Jun. 2011.

[6] A. Morando, T. Victor, and M. Dozza, "A Bayesian reference model for visual time-sharing behaviour in manual and automated naturalistic driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 803–814, Feb. 2020.

[7] J. Wu, Z. Huang, Z. Hu, and C. Lv, "Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving," *Engineering*, early access, Jul. 20, 2022, doi: 10.1016/j.eng.2022.05.017.

[8] T. Horberry et al., "Human-Centered Design for an In-Vehicle Truck Driver Fatigue and Distraction Warning System," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5350–5359, Jun. 2022.

[9] J. H. Hansen, C. Busso, Y. Zheng, and A. Sathyanarayana, "Driver Modeling for Detection and Assessment of Driver Distraction: Examples from the UTDrive Test Bed," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 130–142, Jul. 2017.

[10] X. Zuo, C. Zhang, F. Cong, J. Zhao, and T. Hämäläinen, "Driver Distraction Detection Using Bidirectional Long Short-Term Network Based on Multiscale Entropy of EEG," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 23, 2022, doi: 10.1109/TITS.2022.3159602.

[11] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 300–311, Jun. 2010.

[12] M. Rezaei and R. Klette, "Look at the driver, look at the road: No distraction! No accident!," in *27th Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 129–136.

[13] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton and Y. Zhao, "A Dual-Cameras-Based Driver Gaze Mapping System With an Application on Non-Driving Activities Monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4318–4327, Oct. 2020.

[14] T. K. Chan, C. S. Chin, H. Chen, and X. Zhong, "A Comprehensive Review of Driver Behavior Analysis Utilizing Smartphones," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4444–4475, Oct. 2020.

[15] J. Wang et al., "A Survey on Driver Behavior Analysis From In-Vehicle Cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10186–10209, Aug. 2022.

[16] T. Ersal, H. J. A. Fuller, O. Tsimhoni, J. L. Stein and H. K. Fathy, "Model-Based Analysis and Classification of Driver Distraction Under Secondary Tasks," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 692–701, Sep. 2010.

[17] Y. Xing et al., "Identification and Analysis of Driver Postures for In-Vehicle Driving Activities and Secondary Tasks Recognition," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 1, pp. 95–108, Mar. 2018.

[18] I. Jegham, A. B. Khalifa, I. Alouani, and M.A. Mahjoub,, "A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD," *Signal Process.: Image Commun.*, vol. 88, 115966, Oct. 2020.
[19] A. Aksjonov, P. Nedoma, V. Vodovozov, E. Petlenkov, and M. Herrmann, "Detection and evaluation of driver distraction using machine learning and fuzzy logic," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2048–2059, Jun. 2019.
[20] S. M. Kouchak and A. Gaffar, "Detecting Driver Behavior Using Stacked Long Short Term Memory Network With Attention Layer," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3420–3429, Jun. 2021.
[21] T. H. N. Le, C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "DeepSafeDrive: A grammar-aware driver parsing approach to Driver Behavioral Situational Awareness (DBSAW)," *Pattern Recognit.*, vol. 66, pp. 229–238, Jun. 2017.
[22] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, "Driver Distraction Identification with an Ensemble of Convolutional Neural Networks," *J. Adv. Transp.*, vol. 2019, 4125865, Feb. 2019.
[23] Y. Xing, et al., "Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5379–5390, Jun. 2019.
[24] C. Ryan, F. Murphy, and M. Mullins, "End-to-End Autonomous Driving Risk Analysis: A Behavioural Anomaly Detection Approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1650–1662, Mar. 2021.
[25] C. Ou and F. Karray, "Enhancing Driver Distraction Recognition Using Generative Adversarial Networks," *IEEE Trans. Intell. Veh.*, vol. 5, no. 3, pp. 385–396, Sep. 2020.
[26] B. Baheti, S. Talbar, and S. Gajre, "Towards Computationally Efficient and Realtime Distracted Driver Detection With MobileVGG Network," *IEEE Trans. Intell. Veh.*, vol. 5, no. 4, pp. 565–574, Dec. 2020.
[27] B. Qin, J. Qian, Y. Xin, B. Liu and Y. Dong, "Distracted Driver Detection Based on a CNN With Decreasing Filter Size," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6922–6933, Jul. 2022.
[28] P. Li et al., "Driver Distraction Detection Using Octave-Like Convolutional Neural Network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8823–8833, Jul. 2022.
[29] L. Yang, X. Shan, C. Lv, J. Brighton, and Y. Zhao, "Learning Spatio-Temporal Representations With a Dual-Stream 3-D Residual Network for Nondriving Activity Recognition," *IEEE Trans. Ind. Electron.*, vol. 69, no. 7, pp. 7405–7414, Jul. 2022.
[30] T. Liu, Y. Yang, G. B. Huang, Y. K. Yeo, and Z. Lin, "Driver Distraction Detection Using Semi-Supervised Machine Learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1108–1120, Apr. 2016.
[31] Y. Zhang, Y. Chen, and C. Gao, "Deep unsupervised multi-modal fusion network for detecting driver distraction," *Neurocomputing*, vol. 421, pp. 26–38, Jan. 2021.
[32] B. Li et al., "A New Unsupervised Deep Learning Algorithm for Fine-Grained Detection of Driver Distraction," *IEEE Trans. Intell. Transp. Syst.*, early access, Apr. 18, 2022, doi: 10.1109/TITS.2022.3166275.
[33] C. Huang, P. Hang, Z. Hu, and C. Lv, "Collision-Probability-Aware Human-Machine Cooperative Planning for Safe Automated Driving," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 9752–9763, Oct. 2021.
[34] C. Huang, C. Lv, P. Hang, Z. Hu, and Y. Xing, "Human–Machine Adaptive Shared Control for Safe Driving Under Automation Degradation," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 2, pp. 53–66, Apr. 2022.
[35] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *23th Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 91–100.
[36] Z. Hu, Y. Xing, W. Gu, D. Cao, and C. Lv, "Driver Anomaly Quantification for Intelligent Vehicles: A Contrastive Learning Approach with Representation Clustering," *IEEE Trans. Intell. Veh.*, early access, Mar. 30, 2022, doi: 10.1109/TIV.2022.3163458.
[37] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *18th Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[38] T. Chen et al., "A simple framework for contrastive learning of visual representations," in *37th Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[39] P. Khosla et al., "Supervised Contrastive Learning," in *34th Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.
[40] V. Bazarevsky et al., "BlazePose: On-device Real-time Body Pose tracking," in *36th IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2020.
[41] L. V. D. Maaten, G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, Nov. 2008.