vehicles

MDPI

Article

Highly Discriminative Driver Distraction Detection Method Based on Swin Transformer

Ziyang Zhang [1,2], Lie Yang [1,3,*] and Chen Lv [1,3,*]

[1] State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun 130012, China; 2053741@brunel.ac.uk
[2] Brunel London School, North China University of Technology, Beijing 100144, China
[3] School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798, Singapore
* Correspondence: lie.yang@ntu.edu.sg (L.Y.); lyuchen@ntu.edu.sg (C.L.)

Abstract: Driver distraction detection not only helps to improve road safety and prevent traffic accidents, but also promotes the development of intelligent transportation systems, which is of great significance for creating a safer and more efficient transportation environment. Since deep learning algorithms have very strong feature learning abilities, more and more deep learning-based driver distraction detection methods have emerged in recent years. However, the majority of existing deep learning-based methods are optimized only through the constraint of classification loss, making it difficult to obtain features with high discrimination, so the performance of these methods is very limited. In this paper, to improve the discrimination between features of different classes of samples, we propose a high-discrimination feature learning strategy and design a driver distraction detection model based on Swin Transformer and the highly discriminative feature learning strategy (ST-HDFL). Firstly, the features of input samples are extracted through the powerful feature learning ability of Swin Transformer. Then, the intra-class distance of samples of the same class in the feature space is reduced through the constraint of sample center distance loss (SC loss), and the inter-class distance of samples of different classes is increased through the center vector shift strategy, which can greatly improve the discrimination of different class samples in the feature space. Finally, we have conducted extensive experiments on two publicly available datasets, AUC-DD and State-Farm, to demonstrate the effectiveness of the proposed method. The experimental results show that our method can achieve better performance than many state-of-the-art methods, such as Drive-Net, MobileVGG, Vanilla CNN, and so on.

Keywords: Swin Transformer; driver distraction detection; SC loss; center vector shift; discriminative feature learning

check for updates

Citation: Zhang, Z.; Yang, L.; Lv, C. Highly Discriminative Driver Distraction Detection Method Based on Swin Transformer. *Vehicles* **2024**, *6*, 140–156. https://doi.org/10.3390/vehicles6010006

Academic Editor: David J. Cole

Received: 14 December 2023
Revised: 3 January 2024
Accepted: 9 January 2024
Published: 10 January 2024

1. Introduction

In the increasingly congested and complex road environment, the frequency of traffic accidents and concerns about road safety are constantly escalating. During the driving process, drivers often engage in distraction behaviors such as making phone calls, drinking water, eating, and talking to passengers. These distraction behaviors will lead to slower reaction speed, decreased attention, and decreased perception of the environment, thereby increasing the probability of traffic accidents. Therefore, drivers' distraction often constitutes a major contributing factor to traffic accidents [1]. As a response to enhancing road safety and mitigating accidents, research focused on the detection of drivers' distraction behaviors has garnered significant attention [2]. The research in driver distraction detection aims to develop advanced technological approaches that can monitor and identify deviations from normal driver behavior, enabling timely alerts or warnings to drivers and reducing potential traffic safety risks. This research holds paramount significance in the realms of accident prevention, elevated road safety, and the safeguarding of passengers, pedestrians, and drivers [3]. Moreover, through rigorous data analysis, we can gain a deeper understanding of driver behavior patterns, accident causes, and trends, providing strong support for improving traffic safety policies and driver training. In general, driver distraction detection is of great significance for improving road safety, preventing traffic accidents, and promoting the development of intelligent transportation systems [4]. By reducing driver distraction, we can create a safer, more efficient, and sustainable road transportation environment [5].

Intelligent driving is a driving paradigm that integrates advanced technologies such as artificial intelligence, machine learning, sensor systems, and connectivity to enhance the safety, efficiency, and autonomy of vehicles. It aims to improve the driving experience and reduce human intervention in various aspects of vehicle operation [6]. Driver distraction detection is one of the most important research topics in the field of intelligent driving. In the early stages, researchers have proposed many driver distraction detection methods based on traditional machine learning algorithms to improve road safety [7,8]. These methods are usually trained based on distracted and normal driving behavior data samples, and then use the learned patterns to detect and classify the distraction behavior. These methods usually have good interpretability and can reveal key features that lead to abnormal judgments [9]. Moreover, these methods usually do not require a large amount of data for training, so they are better suited for situations with very limited training data. However, there are also some shortcomings of these traditional machine learning-based driver distraction detection methods. Firstly, the feature selection and extraction process of these methods is relatively difficult, requiring domain expertise, and it is difficult for these methods to fully capture the driver's behavior patterns. In addition, these methods have insufficient generalization performance when dealing with complex and ever-changing driving environments, making it difficult to adapt to complex situations that have not been seen during the training process.

Although traditional machine learning-based methods have certain advantages in driver distraction detection tasks, they also have some limitations in complex scenarios and higher accuracy requirements. With the development of deep learning, many methods based on deep neural networks are constantly emerging to overcome these limitations [10–12]. These deep learning-based methods utilize the powerful feature learning ability of deep neural networks to automatically learn higher-level and abstract feature representations from training data to achieve driver distraction detection with high accuracy. The method based on deep learning can perform end-to-end learning, from raw data to final distraction behavior classification, reducing the need for feature engineering. Moreover, these methods have strong generalization ability and can be applied to various driving scenarios. The application of deep learning methods has greatly promoted the development of the field of driver distraction detection.

These deep learning methods can achieve good performance in driver distraction detection [13]. However, there are still some limitations in these methods. On the one hand, most methods mainly achieve classification through the constraint of cross entropy loss. However, cross entropy loss can only obtain a classification hyperplane that separates samples of different categories, and cannot obtain features with high discrimination [14]. Therefore, the classification accuracy of these methods is difficult to further improve. How to learn features with high discrimination and further improve classification accuracy is a very important challenge currently faced [15]. In addition, the computational complexity of deep learning models is relatively high, requiring a large amount of training data [16]. However, the current dataset in the field of driver distraction detection has a very limited amount of data, making it prone to severe overfitting during the training process. How to overcome overfitting is another important challenge of deep learning-based methods.

To address the aforementioned issues, in this paper, we propose a driver distraction detection method based on Swin Transformer and a highly discriminative feature learning strategy (ST-HDFL). Due to the large receptive field and powerful feature learning abilities of Swin Transformer, it has achieved performance beyond CNN in many studies [17–19].

Therefore, the Swin Transformer is adopted for feature extraction in this paper. However, it is difficult to obtain features with high discrimination only through the constraint of classification loss. Therefore, inspired by [14], a highly discriminative feature learning strategy is proposed in this paper, which includes the constraint of sample and center distance loss (SC loss) and the center vector shift process. Firstly, we initialize a center vector for each class of samples, then reduce the intra-class distance of the same class of samples by minimizing the distance between samples and their corresponding center vectors in the feature space, and improve the inter-class distance of samples of different classes through the center vector shift process. In addition, due to the limited amount of data in existing public datasets related to driver distraction detection, the data augmentation method based on image transformation is adopted to alleviate the overfitting problem. Different from other driver distraction detection methods, our method adopt the powerful feature learning ability of Swin Transformer for feature extraction from the input images, and further improves the discrimination of different class samples in the feature space through the constraint of SC loss and center vector shift strategy, thereby improving the accuracy of driver distraction detection. To evaluate the effectiveness of the proposed driver distraction detection method based on the ST-HDFL model, we have conducted a large number of experiments on the public datasets.

The contributions of this paper are summarized as follows:

1. Due to the powerful image feature learning ability of Swin Transformer, it is introduced to extract more representative features from the input images in this paper.
2. A novel highly discriminative feature learning strategy based on SC loss and center vector shift process is proposed in this paper.
3. To evaluate the effectiveness of the proposed driver distraction detection method, extensive experiments have been conducted on the famous public driver distraction detection datasets (AUC-DDD and State-Farm) in this paper.

The rest of this paper is structured as follows. In Section 2, we reviewed the related works. Section 3 provides a detailed introduction to the proposed driver distraction detection method based on the ST-HDFL model. The experimental dataset, data processing methods, and specific implementation details are introduced in Section 4. Then, the experimental results are processed and analyzed in Section 5. Finally, the research of this paper is summarized and reviewed in Section 6.

2. Related Works

2.1. Driver Distraction Detection Methods

As an important research topic in the field of intelligent driving, driver distraction detection has been widely studied (as shown in Table 1). At first, researchers proposed some methods for detecting driver distraction based on artificial features and machine learning algorithms. For example, support vector machine (SVM) was applied to develop a real-time driver cognitive distraction approach according to drivers' eye movements and their driving performance data in research [20]; in order to estimate the workload of the driver, Zhang et al. [21] proposed a data-driven method based on the decision tree classifier and demonstrated the effectiveness of the method. In [22], Liang et al. developed a driver distraction detection method by combining the dynamic Bayesian network (DBN) and supervised clustering through the analysis of eye movement and driving performance measures of drivers. The training process of these methods was relatively simple, with low requirements for data volume and computing equipment. They were suitable for the situations with limited training data. However, the feature learning ability of the above-mentioned methods is very weak, and the classification ability of the classifier is also very limited. Therefore, the accuracy of these methods is not high enough, and their generalization ability for different scenarios is poor, which cannot meet the needs of practical applications.

Because of the strong feature learning ability and excellent generalization ability, deep neural networks can overcome the shortcomings of traditional machine learning methods and have become more and more popular [16,23,24]. Convolutional neural networks excel at learning high-level abstract features from images, so researchers have proposed a large number of driver distraction detection methods based on convolutional neural networks. For example, Annu Dhiman et al. [25] conducted some comparative analysis of a CNN model with machine learning algorithm and proposed a CNN model that outperforms over VGG16, ResNet50, and logistic regression in the driver distraction detection tasks; to reduce traffic accidents and prevent human lives and property from being damaged, Taimoor Khan et al. [26] developed a convolutional neural network (CNN)-based technique with the integration of a channel attention (CA) mechanism for efficient and effective detection of driver behavior; in [2], the authors proposed a novel lightweight model called multi-stream deep fusion network, which combined transferable CNN features with the high-level semantic feature for the efficient recognition of the driver's state.

Table 1. The related works of driver distraction detection.

References	Description	Methods
[20]	Develop a real-time driver cognitive distraction approach according to drivers' eye movements and their driving performance	SVM
[21]	Propose a data-driven method based on the decision tree classifier	Decision tree
[22]	Develop a driver distraction detection method by combining the dynamic Bayesian network (DBN) and supervised clustering	DBN
[25]	Propose a CNN model that outperforms VGG16, ResNet50, and logistic regression	CNN
[26]	Develop a convolutional neural network (CNN)-based technique with the integration of a channel attention (CA) mechanism	CNN
[2]	Propose a novel lightweight model called multi-stream deep fusion network	CNN

However, the receptive field size of convolutional neural networks is determined by the number and layers of convolutional kernels, and these convolutional neural network-based methods may be limited in capturing features of different scales. The Transformers that have emerged in recent years not only have large receptive fields, but also have powerful feature learning abilities. Therefore, we conduct some exploration on the Transformer-based driver distraction detection method in this paper.

2.2. Research Related to Swin Transformer

Transformer was first proposed by Vaswani et al. in 2017 [27]. It introduced the self-attention mechanism that could effectively capture the dependency relationships between different positions in the input sequence. It was initially widely applied in the field of natural language processing and achieved performance due to other models. The success of Transformer has inspired researchers to explore the application of attention mechanisms in other fields. Due to the local and global relationships of natural images, Dosovitskiy et al. [28] explored the application of transformer to computer vision tasks and proposed the Vision Transformer (ViT) model in 2020. It is renowned for its pure Transformer architecture and has achieved gratifying results in visual tasks [29].

With the proposal of ViT, researchers began to explore the application of Transformer ideas to various computer vision tasks, such as image classification, object detection, image segmentation, and image generation. In addition, various improvements have been made to ViT, including adjusting attention mechanisms, adding multi-scale inputs, etc., to achieve better performance on different tasks. Initially, ViT was mainly applied to image classification tasks. By dividing the image into image blocks and converting the blocks into vector sequences, and then using the self-attention mechanism to capture the relationships between image blocks, ViT has shown comparable or even superior performance in image classification competitions compared to traditional convolutional neural networks (CNNs) [30,31]. The ViT model has also been attempted for target detection tasks. Researchers applied ViT to the field of object detection by dividing the image into grids and treating each grid as an image block, and then using ViT for object detection. This enables ViT to achieve certain success in target detection tasks [32,33]. Some researchers have also explored the application of ViT to image segmentation tasks [34]. By dividing the image into image blocks and using the ViT model on each block, the image can be segmented into different regions for identifying and locating different objects in the image. The ViT model has also been attempted for image generation tasks, such as image generation [35] and super-resolution reconstruction [36]. By training the ViT model to generate different parts of the image, the task of image generation can be achieved.

Although Vision Transformer has strong feature learning ability and broad application prospects, it still has some drawbacks. Firstly, the computational complexity of Vision Transformer is positively correlated with the square of the image size, which makes it difficult to process larger images. In addition, the Vision Transformer is not suitable for tasks where the input image has variable scales. In response to these shortcomings of Vision Transformer, Microsoft researchers have proposed Swin Transformer [37], which is one of the most exciting research developments after the original Vision Transformer. Similar to ViT, Swin Transformer is also widely used for image classification, object detection, image segmentation, and image generation. And it has achieved better performance than the original ViT in many tasks. To make full use of the powerful feature learning abilities of Swin Transformer, we attempt to introduce it into the task of driver distraction detection. Furthermore, we have creatively proposed a highly discriminative feature learning strategy on the basis of Swin Transformer, which is a novel approach to improve the accuracy of driver distraction detection. The research in this paper is beneficial for further improving the accuracy of driving distraction detection, and is of great significance for improving the safety and intelligence level of transportation systems.

3. Methodology

In this paper, we propose a driver distraction detection method based on the ST-HDFL model. The algorithm framework of the ST-HDFL is shown in Figure 1. Firstly, the input image is divided into multiple small patches. Then, each patch is mapped into a feature vector through the linear embedding module. Next, the position bias is added to each feature vector based on the position of each patch in the input image. Then, the Swin Transformer encoder is used to extract features from the feature vector sequence after position encoding. Finally, a fully connected classifier is employed to classify the obtained feature vectors. Before the training process, a center vector is initialized for each class of samples. During the training process, in addition to the constraint of classification loss, we also added the constraint of SC loss to reduce the intra-class distance between the same class of samples. And during each training process, the center vectors are updated through the center vector shift strategy to increase the inter-class distance of different classes of samples in the feature space. In this section, we will provide a detailed introduction to feature extraction through Swin Transformer and the highly discriminative feature learning strategy, respectively.

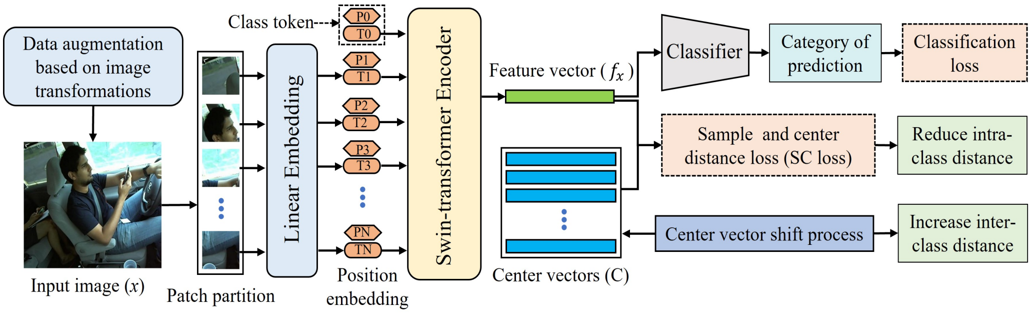


Figure 1. The algorithm framework of the proposed driver distraction detection method based on the ST-HDFL model.

3.1. Feature Extraction through Swin Transformer

Swin Transformer is a deep learning model based on Transformer, which has excellent performance in visual tasks. The overall architecture of the tiny version of Swin Transformer is shown in Figure 2. Different from Vision Transformer, Swin Transformer has the advantages of high accuracy and computational efficiency, and has been used as the backbone of many visual model architectures. Swin Transformer introduces two key concepts to solve the problems faced by the original Vision Transformer: hierarchical feature maps and shifted window attention.

(1) Hierarchical feature maps

The first significant difference from Vision Transformer is that Swin Transformer constructs hierarchical feature maps by gradually merging and downsampling, which allows it to better learn features at different scales. And the non-convolutional downsampling technique patch merging is used in Swin Transformer, which can effectively reduce the dimension of the feature map and reduce the computational complexity.

(2) Shifted window attention

The standard MSA used in Vision Transformer performs global self-attention, and the weight relationship between patches is calculated for all other patches. This leads to the squared complexity of the number of patches, making them unsuitable for high-resolution images. To address this issue, Swin Transformer used a window-based MSA method. A window is just a set of patches, and attention calculation is only performed within each window. Due to the fixed window size throughout the entire network, the complexity of window-based MSA is linearly related to the number of patches, which is a significant improvement over the standard MSA squared complexity.

However, window-based MSA has a significant drawback, which limits self-attention to each window and limits the ability of the network model. To address this issue, Swin Transformer used the shifted window MSA (SW-MSA) module after the W-MSA module. After the shifted operation, a window may consist of non-adjacent patches from the original feature map, so a mask was used in the calculation to limit self-attention to adjacent patches. This shifted window method introduces important cross-connections between windows, which has been proven to improve network performance.

The Swin Transformer replaces the multi-head self-attention (MSA) module of Vision Transformer with window MSA (W-MSA) and shifted window MSA (SW-MSA). The structure of the Swin Transformer block is shown in Figure 2. Each Swin Transformer block consists of two subunits, and each of them consists of a normalization layer, an attention module, followed by another normalization layer and an MLP layer. The first subunit uses the W-MSA module, while the second subunit uses the SW-MSA module. The calculation process for each Swin Transformer block is as follows:

$$\hat{F}_l' = \text{W-MSA}(\text{LN}(F_{l-1})) + F_{l-1}, \tag{1}$$

$$F_l' = \text{MLP}\left(\text{LN}\left(\hat{F}_l'\right)\right) + \hat{F}_l', \tag{2}$$

$$\hat{F}_l = \text{SW-MSA}\left(\text{LN}\left(F_l'\right)\right) + F_l', \tag{3}$$

$$F_l = \text{MLP}(\text{LN}(\hat{F}_l)) + \hat{F}_l \tag{4}$$

where F_{l-1}, $\hat{F}_l'$, F_l', $\hat{F}_l$, and F_l denote the intermediate results of the calculation process, respectively.

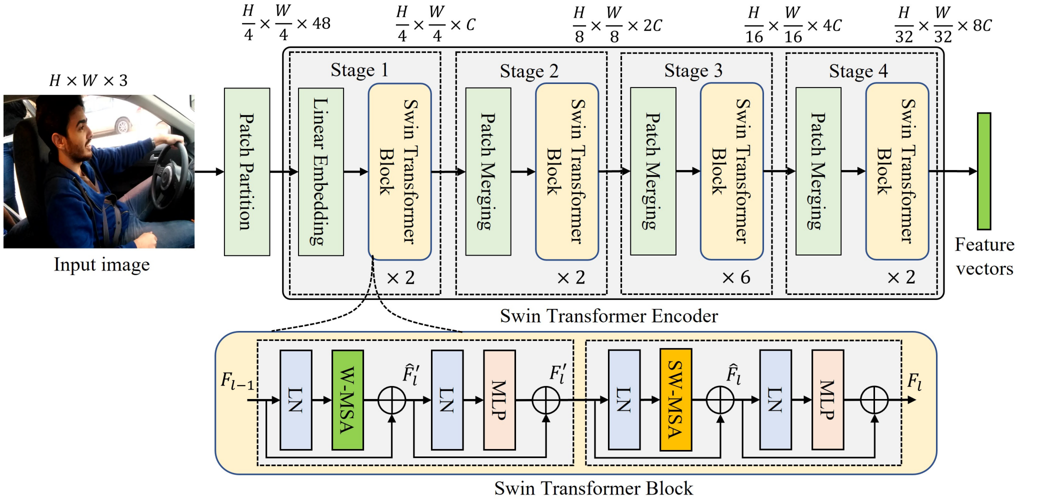


Figure 2. The architecture of the tiny version of Swin Transformer.

Compared with convolutional neural networks, Transformer has the advantages of a large receptive field and high computational efficiency, and has a very broad application prospect in the future. Moreover, in computer vision tasks, Swin Transformer has made some optimizations and improvements on the basis of Vision Transformer, and the feature learning ability and computing efficiency have been significantly improved. Therefore, Swin Transformer is adopted for feature extraction from the input image in the driver distraction detection method proposed in this paper. Firstly, the input images (x) are split into some non-overlapping patches through a patch splitting module. Then, a linear embedding layer is applied to project each patch to a vector $Ti(i = 1, 2, \ldots, N)$, which is treated as a token. Next, in order to fully utilize the relative positional relationship information between different patches, a relative position bias is added to each token as follows:

$$F_0 = [T0; T1; T2; \cdots\cdots; TN] + [P0; P1; P2; \cdots\cdots; PN] \tag{5}$$

where $T0$ is class token, and $Pi(i = 0, 2, \ldots, N)$ denotes the position bias of each token. Finally, several Swin Transformer blocks are applied for feature learning from all the tokens (as shown in Figure 1), and the class token of F_k is taken as the feature vector f_x, where k is the number of Swin Transformer blocks.

3.2. Highly Discriminative Feature Learning Strategy

Optimizing the classification model solely by minimizing classification loss can only obtain a classification hyperplane to divide the input samples into different categories, and cannot obtain features with high discrimination. Therefore, the classification performance is limited. To further improve the accuracy of the proposed driver distraction detection model, a highly discriminative feature learning strategy is proposed in this paper, which includes the center vector initialization process, the constraint of SC loss, and the center vector shift process.

(1) Center vector initialization process.

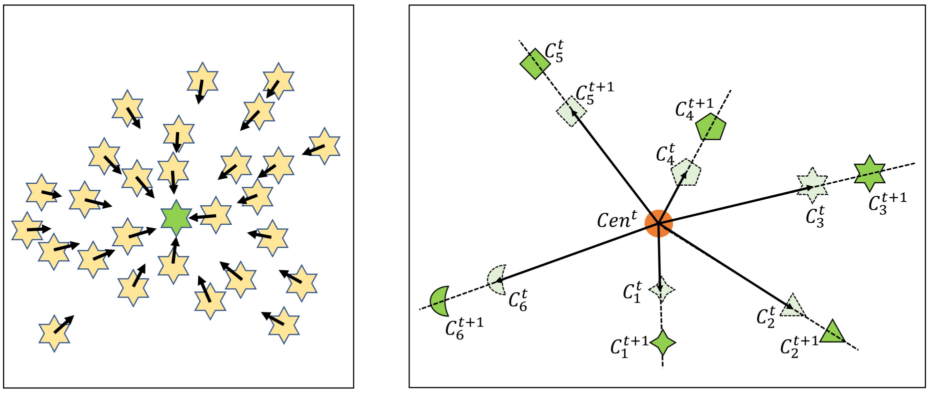
Before the training process, a center vector needs to be randomly initialized for each class of samples. During the initialization process, the number of center vectors is the same as the number of sample categories in the experimental datasets, and the dimensions of each center vector are consistent with the feature vectors extracted by Swin Transformer.

(2) The constraint of SC loss.

In order to reduce the intra-class distance between the same classes of samples in the feature space, the constraint of SC loss is introduced into the proposed ST-HDFL model. The SC loss is the average distance between the feature vectors and their corresponding center vectors. During the training process, we promote the aggregation of the same class of samples to their corresponding center vectors through minimizing the SC loss (as shown in Figure 3a), which can greatly reduce the intra-class distance of samples of the same class in the feature space. The calculation formula of SC loss is as follows:

$$L_{sc} = \frac{1}{B}\sum_{i=1}^{B}\left\| f_i^t - C_{y_i}^t \right\|_2 \tag{6}$$

where f_i^t denotes the feature vector of the i-th sample in the t-th iteration; B denotes the number of samples in a batch; y_i is the label of the i-th sample in the training batch; $C_{y_i}^t$ represents the central vector of the y_i-th class of samples during the t-th iteration.



(a) The constraint of SC loss

(b) The center vector shift process

Figure 3. The schematic diagram of the highly discriminative feature learning strategy. In (**a**), the yellow hexagons represent the feature vectors of one class of samples, and the green hexagon represents the corresponding center vector of this class of samples; in (**b**), Cen^t is the average vector of all central vectors during the t-th iteration; C_i^t and C_i^{t+1} are the central vectors of samples with the label of i before and after the central vector shift process in the t-th iteration.

(3) Center vector shift process.

Although the constraint of SC loss can promote each class of samples to cluster near the corresponding center vector, due to the insufficient distance between different center vectors, the discrimination among different classes of samples is usually not obvious. To improve the inter-class distance of different classes of samples in the feature space, the center vector shift process is adopted to the proposed ST-HDFL model. During the center vector shift process, we need to calculate the average vector Cen of all center vectors at first, then let each center vector shift along the direction of $\overrightarrow{Cen, C}$ by a certain step size (as shown in Figure 3b). The mathematical formula of the central vector shift process can be described as

$$Cen^t = \frac{1}{n_c} \sum_{i=1}^{n_c} C_i^t \tag{7}$$

$$C_i^{t+1} = C_i^t + \alpha \cdot \frac{\left(C_{y_i}^t - Cen^t\right)}{\left\| C_{y_i}^t - Cen^t \right\|_2} \tag{8}$$

where Cen^t is the average vector of all central vectors during the t-th iteration; n_c is the number of sample categories in the experimental datasets; α is the step size of the central vector shift process; C_i^t and C_i^{t+1} are the central vectors of samples with the label of i before and after the central vector shift process in the t-th iteration, respectively.

4. Experimental Datasets and Experiment Setup

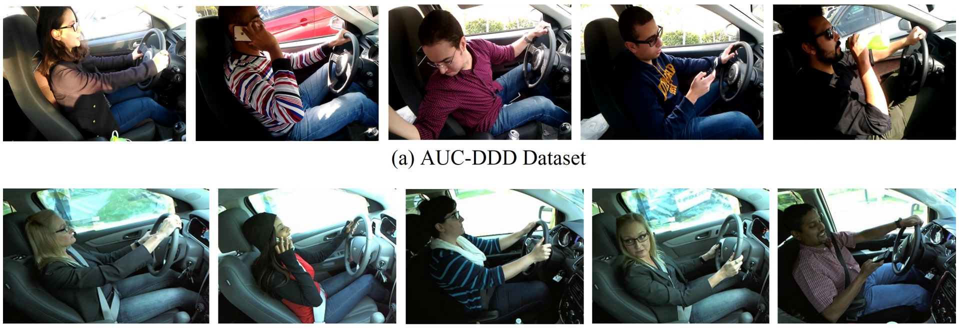
4.1. Experimental Datasets

(1) AUC-DDD dataset

AUC Distracted Driver Detection (AUC-DDD) dataset is one of the most famous driver distraction datasets; it was created by Abouelnaga et al. [38] at the American University in Cairo. There are a total of 17,308 images of 31 participants from seven different countries in this dataset; these images were randomly split into the training set (including 12,977 images) and validation set (including 4331 images). All samples in this dataset were collected on real vehicles. In order to ensure safety, the vehicles were parked in a safe area along the road during the data collection process. And the size of each image in this dataset was 1920 × 1080. Some samples of this dataset are shown in Figure 4a. All samples in this dataset are divided into 10 classes: safe driving, talking on the phone with the left or right hand, texting with the left or right hand, eating or drinking, reaching behind, fixing hair and makeup, adjusting the radio, and talking to passengers.

(2) State-Farm dataset

The State Farm Distracted Driver Detection (State-Farm) dataset, provided by the Kaggle competition, includes 22,424 labeled images and 79,726 unlabeled images. All the labeled images are divided into 10 classes, and the classification of these samples is the same as that of the AUC dataset. The size of all sample images in this dataset was 640 × 480. Similar to the AUC-DDD dataset, all samples in this dataset were collected from real vehicles parked in safe areas along the road. In the experiments of this paper, we randomly select 80% of samples as the training set, and the remaining 20% of samples are taken as the test set. Some labeled samples of this dataset are shown in Figure 4b.



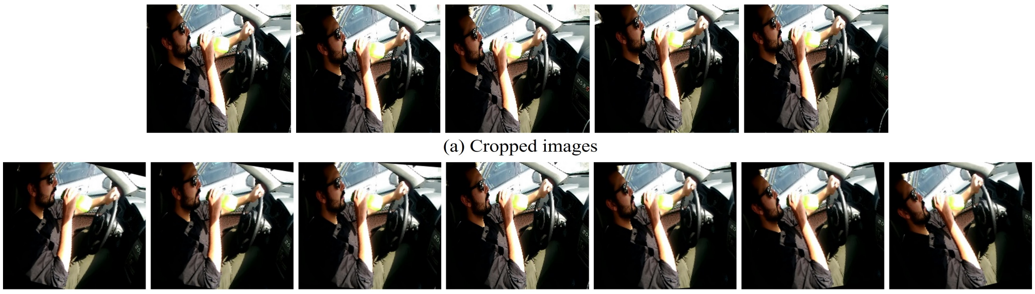
(a) AUC-DDD Dataset

(b) State-Farm Dataset

Figure 4. Some samples of the experimental datasets.

4.2. Data Augmentation Based on Image Transformation

To alleviate overfitting, inspired by [23], some data augmentation methods based on image transformation are adopted in this paper to obtain more training data. Firstly, all samples are resized to 250 × 250, then we select (125, 125), (112, 112), (112, 138), (138, 112), (138, 138) as the central points, respectively, and crop out the patches with the size of 224 × 224 from each sample. Some cropped images obtained from one sample are shown in Figure 5a. Next, the cropped images are rotated by -15°, -10°, -5°, 0°, 5°, 10°, and 15°, respectively. Some images rotated from one cropped image are shown in Figure 5b. After data augmentation, the training sets of each experimental dataset are 35 times (35 = 5 × 7) larger than the original one.

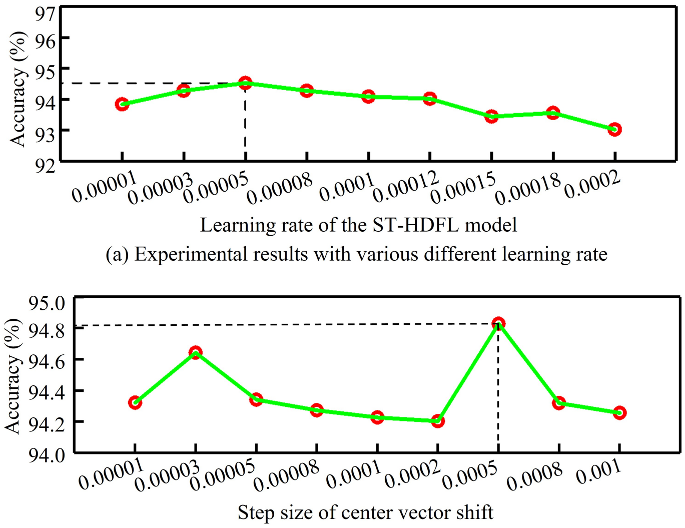

(a) Cropped images

(b) Images rotated from the same sample

Figure 5. Some examples of data augmentation based on image transformation.

4.3. Implementation Details

Due to the limited amount of data in the experimental data, in order to overcome the overfitting problem, we selected the tiny version of Swin Transformer with the least number of parameters in our experiment, and loaded the official pre training model before the training process. During the training process, the learning rate of model optimization is an important parameter that affects the performance of the model. So, some experiments are conducted for learning rate selection on the AUC-DDD dataset, and the experimental results are shown in Figure 6a. According to the figure, the learning rate is set as $lr = 0.00005$ in the following experiments of this paper. In addition, the step size of the center vector shift process is a very important parameter for the highly discriminative feature learning strategy. To find an appropriate step size for the center vector shift process, we conduct some experiments on the AUC dataset with multiple different step sizes. According to the experimental results (as shown in Figure 6b), the step size of the center vector shift process is set as $\alpha = 0.0005$ in the following experiments. The Adam optimizer is selected to optimize the proposed model, and the parameters of the optimizer are set as $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Moreover, the batch size is set to 24, and the weight of SC loss in the full objective function is set as $\lambda = 0.95$. After training for 50 epochs, the network can converge well, and the experimental results in this paper are all obtained by training for 50 epochs. In order to speed up the training process, the proposed mode is implemented with PyTorch on the platform with two Nvidia GeForce RTX 3090 graphic cards.



(a) Experimental results with various different learning rate

(b) Experimental results with different step size of center vector shift

Figure 6. The results of some key parameter selection experiments on the AUC-DDD dataset.

5. Experimental Results

In order to demonstrate the effectiveness of data augmentation and the driver distraction detection method based on the ST-HDFL model proposed in this paper, a large number of experiments were conducted on the AUC-DDD dataset and State-Farm dataset, respectively. And all the experimental results are processed and analyzed in this section.

5.1. Evaluation of the Highly Discriminative Feature Learning Strategy

In this paper, we propose a highly discriminative feature learning strategy that reduces the intra-class distance of samples of the same class in the feature space through the SC loss constraint, and improves the inter-class distance of samples of different classes through the center vector shift process. In this section, we have conducted some experiments to demonstrate that this strategy can improve the discrimination of different class samples in the feature space and improve the accuracy of the proposed driver distraction detection model.

To evaluate the effectiveness of the SC loss constraint and center vector shift process of the proposed highly discriminative feature learning strategy, some contrast experiments are conducted on the AUC-DDD dataset. Firstly, we conduct the experiment without adopting the SC loss and center vector shift process, and take it as the SC-Exp. Then, we introduce the SC loss on the basis of the SC-Exp and take it as Shift-Exp. Next, the experiment introducing the center vector shift process on the basis of the Shift-Exp is the experiment of our method. To better reflect the differences between these experimental results, we have conducted 10 runs for SC-Exp, Shift-Exp, and our method, respectively. And the accuracy and average accuracy of these 10 runs are shown in Table 2. Additionally, we performed paired T-tests on the experimental results between SC-Exp and Shift-Exp, as well as between Shift-Exp and our method, and the results are shown in Table 3. According to Tables 2 and 3, the experimental results of Shift-Exp are significantly better than those of SC-Exp, which proves that the constraint of SC loss is beneficial for improving the accuracy of the proposed model. Although the variance of 10 runs slightly increases after introducing the center vector shift process on the basis of Shift-Exp, the average accuracy of our method is higher than that of Shift-Exp and there is a significant difference between the results of our method and Shift-Exp (as shown in the two tables), which confirms the effectiveness of the center vector shift process.

Table 2. Results of the highly discriminative feature learning strategy evaluation experiments. In this table, bold data represent the best values, while underlined data represent second-best values.

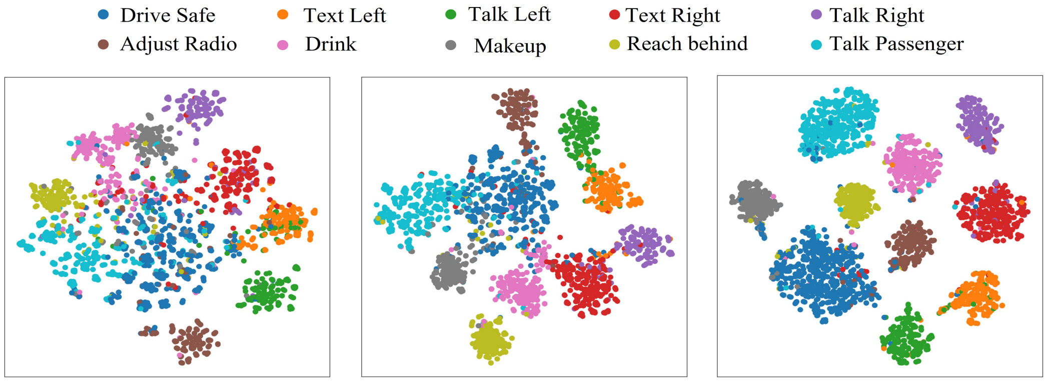
Exps	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Aver Acc ↑**	**STD ↓**
SC-Exp	94.26	94.87	94.96	95.27	95.13	94.91	94.79	94.95	94.89	94.78	94.881	0.264
Shift-Exp	95.21	95.17	95.45	94.96	95.39	95.27	95.63	95.46	95.32	95.43	95.329	**0.187**
Our method	95.33	95.62	95.78	95.81	95.63	95.21	95.97	95.75	95.63	95.58	**95.631**	0.224

Table 3. The paired T-test results based on Table 1.

Paired T-test	Shift-Exp vs. SC-Exp	Our method vs. Shift-Exp
p-value	†† (0.000367)	† (0.00429)

Note: ∼nonsignificant, * ($p \leq 0.05$), ** ($p \leq 0.01$), † ($p \leq 0.005$), †† ($p \leq 0.001$).

To further demonstrate that the constraint of SC loss can make the same class of samples gather more tightly in the feature space, and that the center vector shift strategy can increase the distance of different classes of samples, we separately map the feature vectors of all samples obtained from SC-Exp, Shift-Exp, and our method to a two-dimensional plane through the T-SNE algorithm [39] (as shown in Figure 7). In the SC-Exp, the model is optimized only relying on the constraint of classification loss. Although the features of different classes of samples tend to converge to the same region, the discrimination of different classes of samples in the feature space is not significant enough (as shown in Figure 7a). It can be seen from Figure 7b that after introducing the constraint of SC loss, samples of the same class gather more closely in the feature space. The results prove that SC loss can effectively reduce the distance between samples of the same class in the feature space. However, there are still some overlapping areas in the feature space for samples of different classes (as shown in Figure 7b). After introducing the center vector shift process, the distance between samples of different classes in the feature space significantly increases (as shown in Figure 7c), which demonstrates that the center vector shift process can effectively improve the discrimination of different classes of samples. After sufficient training, the method proposed in this paper can extract features with high discrimination; therefore, it can achieve very excellent performance in driver distraction detection tasks.



(a) Baseline (b) SC Experiment (c) Shift Experiment

Figure 7. The feature distribution of the highly discriminative feature learning strategy evaluation experiments.

Due to the proposal of hierarchical feature maps and shifted window attention, Swin Transformer not only has high computational efficiency but also can fully learn multi-scale features, so it has quite excellent feature learning ability. To demonstrate its superiority in feature learning, some comparison experiments between Swin Transformer and five famous classification models, AlexNet [40], VGG [41], ResNet [42], MobileNet [43], and ViT [28], respectively, were conducted. To alleviate the over-fitting problem, the official pre-training model is loaded before the model training process, and then we fine-tune each model on the AUC-DDD dataset, respectively. All the experimental results are shown in Figure 8. According to the figure, the precision, recall, F1 score, and accuracy of Swin Transformer are better than those of all the classical models involved in the comparison, which proves that Swin Transformer has better feature learning ability than other classical models in the driver distraction detection task.

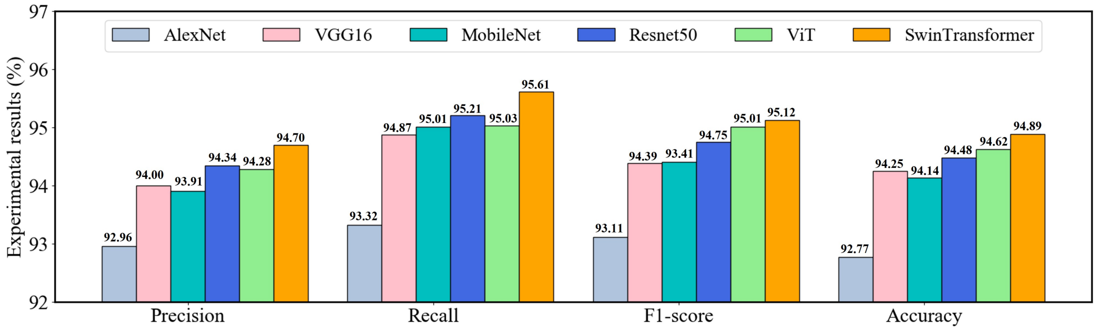


Figure 8. The experimental result comparison between our ST-HDFL model and five famous classification models on the AUC-DDD dataset.

5.2. Comparison Experiments on the Public Datasets

To further evaluate the performance of our method, we have conducted some experiments on two public driver distraction detection datasets AUC-DDD and State-Farm, and compare the experimental results of our method and other state-of-the-art methods, as shown in Tables 4 and 5, respectively. It can be seen from Table 4 that the accuracy of the proposed ST-HDFL-based method is higher than that of all the state-of-the-art methods involved in the comparison on the AUC-DDD dataset. Moreover, according to Table 5, our method can also achieve better performance than all the state-of-the-art methods participating in the comparison except for D-HCNN [44] and CAT-CapsNet [45] on the State-Farm dataset. Overall, compared to the current state-of-the-art methods, our method can achieve very excellent performance on both the AUC-DDD dataset and State-Farm dataset.

Table 4. The comparison of experimental results between the proposed ST-HDFL model and some state-of-the-art methods on the AUC-DDD dataset.

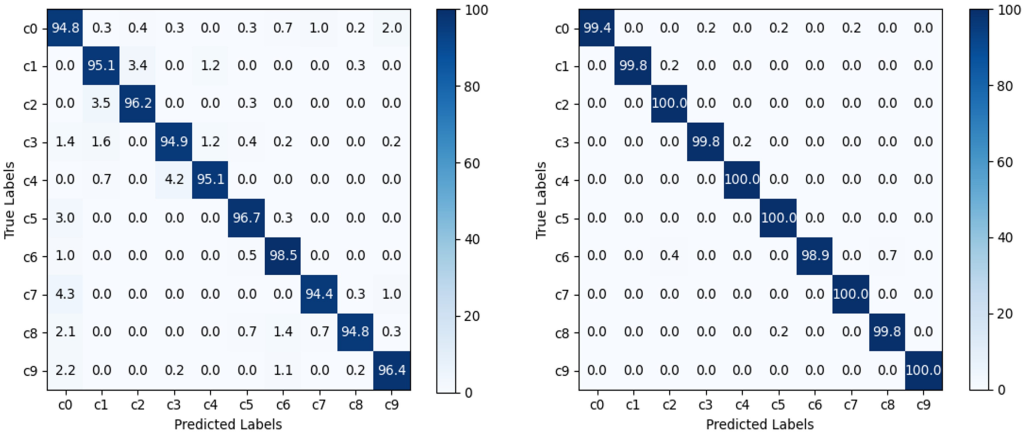
Experiments	Accuracy (%)
ADNet [46]	90.22
BiRSwinT [47]	92.25
NasNet Mobile [48]	94.69
FRNet [49]	94.74
MobileVGG [12]	95.25
D-HCNN [44]	95.59
ST-HDFL (ours)	95.66

Table 5. The comparison of experimental results between the proposed ST-HDFL model and some state-of-the-art methods on the State-Farm dataset.

Experiments	Accuracy (%)
VGG16 + VGG-GAP [50]	92.60
Drive-Net [51]	95.00
HCF [52]	96.74
Vanilla CNN [53]	97.05
D-HCNN [44]	99.86
CAT-CapsNet [45]	99.88
ST-HDFL (ours)	99.73

5.3. Discussion

The two experimental datasets, AUC-DDD and State-Farm, adopted in our study both consist of 10 classes of samples. To further explore the classification performance of the proposed model on these ten class samples, we have drawn the confusion matrix of the experimental results on the two experimental datasets (as shown in Figure 9). According to the figure, the proposed ST-HDFL model exhibits relatively uniform classification accuracy across the ten classes, and the accuracy of each class is above 90%, which further demonstrates the excellent classification performance of our method.



	c0	c1	c2	c3	c4	c5	c6	c7	c8	c9
c0	94.8	0.3	0.4	0.3	0.0	0.3	0.7	1.0	0.2	2.0
c1	0.0	95.1	3.4	0.0	1.2	0.0	0.0	0.0	0.3	0.0
c2	0.0	3.5	96.2	0.0	0.0	0.3	0.0	0.0	0.0	0.0
c3	1.4	1.6	0.0	94.9	1.2	0.4	0.2	0.0	0.0	0.2
c4	0.0	0.7	0.0	4.2	95.1	0.0	0.0	0.0	0.0	0.0
c5	3.0	0.0	0.0	0.0	0.0	96.7	0.3	0.0	0.0	0.0
c6	1.0	0.0	0.0	0.0	0.0	0.5	98.5	0.0	0.0	0.0
c7	4.3	0.0	0.0	0.0	0.0	0.0	0.0	94.4	0.3	1.0
c8	2.1	0.0	0.0	0.0	0.0	0.7	1.4	0.7	94.8	0.3
c9	2.2	0.0	0.0	0.2	0.0	0.0	1.1	0.0	0.2	96.4

(a) Confusion matrix on AUC-DDD dataset

	c0	c1	c2	c3	c4	c5	c6	c7	c8	c9
c0	99.4	0.0	0.0	0.2	0.0	0.2	0.0	0.2	0.0	0.0
c1	0.0	99.8	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
c2	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
c3	0.0	0.0	0.0	99.8	0.2	0.0	0.0	0.0	0.0	0.0
c4	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
c5	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
c6	0.0	0.0	0.4	0.0	0.0	0.0	98.9	0.0	0.7	0.0
c7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
c8	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	99.8	0.0
c9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

(b) Confusion matrix on State-Farm dataset

Figure 9. The confusion matrices of the experimental results on AUC-DDD dataset and State-Farm dataset. In the figure, c0 c9 represent safe driving, talking on the phone with the left hand, talking on the phone with the right hand, texting with the left hand, texting with the right hand, eating or drinking, reaching behind, fixing hair and makeup, adjusting the radio, and talking to passengers, respectively.

Although the proposed model ST-HDFL can achieve very excellent performance on the driver distraction detection task, there are also some limitations of this method. Firstly, compared with some classical convolutional models, such as AlexNet, VGG16, MobileNet, and ResNet50, Swin Transformer has higher computational complexity, so it is more time-consuming during the calculation process. As depicted in Table 6, it is evident that the average processing time for a single image by the Swin Transformer is significantly higher than that of other classical models. In addition, due to the fact that the distraction behavior of the driver is a continuous process, and this model is targeted at a single image, it cannot fully utilize the temporal features between adjacent frames, so the performance of this model is very limited. Therefore, in future research, we will attempt to develop lighter models while ensuring accuracy. In addition, to make full use of the continuous temporal features between adjacent frames, we will attempt to design a video-based driver distraction detection method.

Table 6. The average running time required for each model to process an image during the test phase.

Methods	AlexNet	VGG16	MobileNet	ResNet 50	Swin Transformer
Time cost	0.52 ms	1.13 ms	2.55 ms	2.85 ms	4.81 ms

6. Conclusions

In this paper, the ST-HDFL model is designed for driver distraction detection, which can efficiently extract features from input images through Swin Transformer, and then further improve the discrimination of different classes of samples in the feature space through the proposed highly discrimination feature learning strategy, thereby achieving high accuracy in driver distraction behavior recognition tasks. In addition, extensive experiments on the public datasets have verified the effectiveness of the proposed highly discrimination feature learning strategy and excellent feature learning ability of Swin Transformer. According to the experimental results, our driver distraction detection method achieved an accuracy of 95.66% on the famous public dataset AUC-DDD and 99.71% on the State-Farm dataset. Some comparison experiments on two public datasets have proved that the proposed ST-HDFL model can achieve better performance than many state-of-the-art methods. In future applications, the proposed method can accurately detect the distraction behavior of drivers and provide timely reminders to drivers to minimize the occurrence of some traffic accidents. Overall, the research in this paper is of great significance for improving driving safety.

Author Contributions: Conceptualization, Z.Z. and L.Y.; methodology, Z.Z. and L.Y.; software, Z.Z. and L.Y.; validation, Z.Z. and L.Y.; formal analysis, Z.Z. and L.Y.; investigation, Z.Z. and L.Y.; resources, L.Y. and C.L.; data curation, Z.Z. and L.Y.; writing—original draft preparation, Z.Z. and L.Y.; writing—review and editing, Z.Z., L.Y. and C.L.; visualization, L.Y. and C.L.; supervision, C.L.; project administration, C.L.; funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Jilin University: Foundation of State Key Laboratory of Automotive Simulation and Control.

Data Availability Statement: The dataset used in this article is the well-known public dataset AUC-DDD and State-Farm, and we have not created a new dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, J.; Chai, W.; Venkatachalapathy, A.; Tan, K.L.; Haghighat, A.; Velipasalar, S.; Adu-Gyamfi, Y.; Sharma, A. A survey on driver behavior analysis from in-vehicle cameras. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 10186–10209. [CrossRef]
2. Hu, Z.; Xing, Y.; Gu, W.; Cao, D.; Lv, C. Driver anomaly quantification for intelligent vehicles: A contrastive learning approach with representation clustering. *IEEE Trans. Intell. Veh.* **2022**, *8*, 37–47. [CrossRef]
3. Tan, M.; Ni, G.; Liu, X.; Zhang, S.; Wu, X.; Wang, Y.; Zeng, R. Bidirectional posture-appearance interaction network for driver behavior recognition. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 13242–13254. [CrossRef]
4. Kashevnik, A.; Shchedrin, R.; Kaiser, C.; Stocker, A. Driver distraction detection methods: A literature review and framework. *IEEE Access* **2021**, *9*, 60063–60076. [CrossRef]
5. Alemdar, K.D.; Kayacı Çodur, M.; Codur, M.Y.; Uysal, F. Environmental Effects of Driver Distraction at Traffic Lights: Mobile Phone Use. *Sustainability* **2023**, *15*, 15056. [CrossRef]
6. Meiring, G.A.M.; Myburgh, H.C. A review of intelligent driving style analysis systems and related artificial intelligence algorithms. *Sensors* **2015**, *15*, 30653–30682. [CrossRef]
7. Zhao, C.; Zhang, B.; He, J.; Lian, J. Recognition of driving postures by contourlet transform and random forests. *IET Intell. Transp. Syst.* **2012**, *6*, 161–168. [CrossRef]
8. Zhang, X.; Zheng, N.; Wang, F.; He, Y. Visual recognition of driver hand-held cell phone use based on hidden CRF. In Proceedings of the 2011 IEEE International Conference on Vehicular Electronics and Safety, Beijing, China, 10–12 July 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 248–251
9. Feng, S.; Yan, X.; Sun, H.; Feng, Y.; Liu, H.X. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nat. Commun.* **2021**, *12*, 748. [CrossRef]
10. Yang, H.; Wu, J.; Hu, Z.; Lv, C. Real-Time Driver Cognitive Workload Recognition: Attention-Enabled Learning with Multimodal Information Fusion. *IEEE Trans. Ind. Electron.* **2023**, *71*, 4999–5009. [CrossRef]
11. Yang, H.; Liu, H.; Hu, Z.; Nguyen, A.-T.; Guerra, T.-M.; Lv, C. Quantitative Identification of Driver Distraction: A Weakly Supervised Contrastive Learning Approach. *IIEEE Trans. Intell. Transp. Syst.* 2023, *early access*. [CrossRef]
12. He, X.; Wu, J.; Huang, Z.; Hu, Z.; Wang, J.; Sangiovanni-Vincentelli, A.; Lv, C. Fear-Neuro-Inspired Reinforcement Learning for Safe Autonomous Driving. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 267–279. [CrossRef]
13. Mase, J.M.; Chapman, P.; Figueredo, G.P.; Torres, M.T. A hybrid deep learning approach for driver distraction detection. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 21–23 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
14. Yang, L.; Song, Y.; Ma, K.; Xie, L. Motor imagery EEG decoding method based on a discriminative feature learning strategy. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 368–379. [CrossRef] [PubMed]
15. Yang, L.; Yang, H.; Hu, B.-B.; Wang, Y.; Lv, C. A Robust Driver Emotion Recognition Method Based on High-Purity Feature Separation. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 15092–15104. [CrossRef]

16. Yang, L.; Song, Y.; Ma, K.; Su, E.; Xie, L. A novel motor imagery EEG decoding method based on feature separation. *J. Neural Eng.* **2021**, *18*, 036022. [CrossRef]
17. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.-S.; Khan, F.S. Transformers in Remote Sensing: A Survey. *Remote Sens.* **2023**, *15*, 1860. [CrossRef]
18. Chen, T.; Mo, L. Swin-fusion: Swin-transformer with feature fusion for human action recognition. *Neural Process. Lett.* **2023**, *55*, 11109–11130. [CrossRef]
19. Xiao, H.; Li, L.; Liu, Q.; Zhu, X.; Zhang, Q. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control* **2023**, *84*, 104791. [CrossRef]
20. Liang, Y.; Reyes, M.L.; Lee, J.D. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Trans. Intell. Transp. Syst.* **2007**, *8*, 340–350. [CrossRef]
21. Zhang, Y.; Owechko, Y.; Zhang, J. Driver cognitive workload estimation: A data-driven perspective. In Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749), Washington, WA, USA, 3–6 October 2004; IEEE: Piscataway, NJ, USA, 2004; pp. 642–647.
22. Liang, Y.; Lee, J.D.; Reyes, M.L. Nonintrusive detection of driver cognitive distraction in real time using Bayesian networks. *Transp. Res. Board* **2007**, *2018*, 1–8. [CrossRef]
23. Yang, L.; Tian, Y.; Song, Y.; Yang, N.; Ma, K.; Xie, L. A novel feature separation model exchange-GAN for facial expression recognition. *Knowl.-Based Syst.* **2020**, *204*, 106217. [CrossRef]
24. Guo, Z.; You, L.; Liu, S.; He, J.; Zuo, B. ICMFed: An Incremental and Cost-Efficient Mechanism of Federated Meta-Learning for Driver Distraction Detection. *Mathematics* **2023**, *11*, 1867. [CrossRef]
25. Dhiman, A.; Varshney, A.; Hasani, F.; Verma, B. A Comparative Study on Distracted Driver Detection Using CNN and ML Algorithms. In Proceedings of the International Conference on Data Science and Applications, London, UK, 25–26 November 2023; Lecture Notes in Networks and Systems; Saraswat, M., Chowdhury, C., Kumar Mandal, C., Gandomi, A.H., Eds.; Springer: Singapore, 2023; Volume 552.
26. Khan, T.; Choi, G.; Lee, S. EFFNet-CA: An efficient driver distraction detection based on multiscale features extractions and channel attention mechanism. *Sensors* **2023**, *23*, 3835. [CrossRef] [PubMed]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
29. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
30. Ma, Y.; Wang, Z. ViT-DD: Multi-Task Vision Transformer for Semi-Supervised Driver Distraction Detection. In Proceedings of the IEEE Intelligent Vehicles Symposium Workshops (IV Workshops), Anchorage, AK, USA, 4 June 2023.
31. Peng, K.; Roitberg, A.; Yang, K.; Zhang, J.; Stiefelhagen, R. TransDARC: Transformer-based driver activity recognition with latent space feature calibration. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 278–285
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
33. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
34. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
35. Jiang, Y.; Chang, S.; Wang, Z. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 14745–14758.
36. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12873–12883.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
38. Abouelnaga, Y.; Eraqi, H.M.; Moustafa, M.N. Real-time distracted driver posture classification. *arXiv* **2017**, arXiv:1706.09498.
39. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
40. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [CrossRef]
41. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

43. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
44. Qin, B.; Qian, J.; Xin, Y.; Liu, B.; Dong, Y. Distracted driver detection based on a CNN with decreasing filter size. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 6922–6933. [CrossRef]
45. Mittal, H.; Verma, B. CAT-CapsNet: A Convolutional and Attention Based Capsule Network to Detect the Driver's Distraction. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9561–9570. [CrossRef]
46. Xiao, W.; Liu, H.; Ma, Z.; Chen, W. Attention-based deep neural network for driver behavior recognition. *Futur. Gener. Comput. Syst.* **2022**, *132*, 152–161. [CrossRef]
47. Yang, W.; Tan, C.; Chen, Y.; Xia, H.; Tang, X.; Cao, Y.; Zhou, W.; Lin, L.; Dai, G. BiRSwinT: Bilinear full-scale residual swin-transformer for fine-grained driver behavior recognition. *J. Frankl. Inst.* **2023**, *360*, 1166–1183. [CrossRef]
48. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8697–8710.
49. Duan, C.; Gong, Y.; Liao, J.; Zhang, M.; Cao, L. FRNet: DCNN for Real-Time Distracted Driving Detection toward Embedded Deployment. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9835–9848. [CrossRef]
50. Zhang, B. *Apply and Compare Different Classical Image Classification Method: Detect Distracted Driver*; Computer Science Department, Stanford University: Stanford, CA, USA, 2016.
51. Majdi, M.S.; Ram, S.; Gill, J.T.; Rodríguez, J.J. Drive-net: Convolutional network for driver distraction detection. In Proceedings of the 2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), Las Vegas, NV, USA, 8–10 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
52. Huang, C.; Wang, X.; Cao, J.; Wang, S.; Zhang, Y. HCF: A hybrid CNN framework for behavior detection of distracted drivers. *IEEE Access* **2020**, *8*, 109335–109349. [CrossRef]
53. Janet, B.; Reddy, U.S. Real time detection of driver distraction using CNN. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 185–191.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.