

A Novel Driver Distraction Behavior Detection Method Based on Self-Supervised Learning With Masked Image Modeling

Yingzhi Zhang[✉], Graduate Student Member, IEEE, Taiguo Li[✉], Chao Li, and Xinghong Zhou

Abstract—Driver distraction causes a significant number of traffic accidents every year, resulting in economic losses and casualties. Currently, the level of automation in commercial vehicles is far from completely unmanned, and drivers still play an important role in operating and controlling the vehicle. Therefore, driver distraction behavior detection is crucial for road safety. Presently, driver distraction detection primarily relies on traditional convolutional neural networks (CNNs) and supervised learning methods. However, there are still challenges, such as the high cost of labeled data sets, limited ability to capture high-level semantic information, and weak generalization performance. In order to solve these problems, this article proposes a new self-supervised learning method based on masked image modeling (MIM) for driver distraction behavior detection. First, a self-supervised learning framework for MIM is introduced to solve the serious human and material consumption issues caused by data set labeling. Second, the Swin Transformer is employed as an encoder. Performance is enhanced by reconfiguring the Swin Transformer block and adjusting the distribution of the number of window multihead self-attention (W-MSA) and shifted W-MSA (SW-MSA) detection heads across all stages, which leads to model more lightening. Finally, various data augmentation strategies are used along with the best random masking strategy to strengthen the model's recognition and generalization ability. Test results on a large-scale driver distraction behavior data set show that the self-supervised learning method proposed in this article achieves an accuracy of 99.60%, approximating the excellent performance of advanced supervised learning methods. Our code is publicly available at github.com/Rocky1salady-killer/SL-DDBD.

Index Terms—Driver distraction, masked image modeling (MIM), self-supervised learning, vision transformer (ViT).

I. INTRODUCTION

RIVER behavior and decision control of vehicles are the main factors affecting safe driving. According to a report published by the World Health Organization (WHO),

Manuscript received 13 July 2023; accepted 22 August 2023. Date of publication 28 August 2023; date of current version 6 February 2024. This work was supported in part by the Science and Technology Program of Gansu Province under Grant 21JR7RA303; in part by the Gansu Provincial Department of Education: University Teacher Innovation Fund Project under Grant 2023A-039; and in part by the Lanzhou Jiaotong University Youth Science Foundation Project under Grant 2020002. (*Corresponding author: Taiguo Li*)

Yingzhi Zhang, Taiguo Li, and Xinghong Zhou are with the College of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China (e-mail: 12211496@stu.lzjtu.edu.cn; leetg@mail.lzjtu.cn; 12221581@stu.lzjtu.edu.cn).

Chao Li is with the Children's Rehabilitation Department, Shaanxi Kangfu Hospital, Xi'an 710065, China (e-mail: 544738669@qq.com).

Digital Object Identifier 10.1109/JIOT.2023.3308921

approximately 1.35 million people worldwide are killed in traffic accidents each year. Between 20 and 50 million people are injured and become disabled because of traffic accidents [1]. Studies also indicate that driver distractions are one of the important causes of road traffic accidents. Additionally, according to data from the National Highway Traffic Safety Administration (NHTSA) [2], in USA, driver distractions contributed to 3142 traffic accidents in 2019. The American Society of Automotive Engineers (SAEs) divides autonomous driving into six levels, ranging from L0 to L5. By 2030, USA, Europe, and China will have 80 million L4/L5 intelligent vehicles. Although autonomous driving technology has made impressive progress, the control of vehicles by autonomous driving is still immature [3]. In the autonomous driving tests conducted by Uber, 37 traffic accidents were related to driver distraction. The driver's distracting behavior caused the vehicle to not take control in emergency situations and implement emergency remedial actions in a timely manner [4]. Therefore, whether it is autonomous driving or manual driving, the driver needs to remain focused during the vehicle's journey. The driver's state is particularly important, and an efficient and accurate driver distraction detection system is an important research method for achieving traffic safety [5]. Driver distraction detection will be integrated into Advanced Driver Assistance Systems (ADASs), analyzing the driver's actions and behavior to predict unsafe distraction operations [6]. When the driver's distraction is detected, the vehicle dashboard displays a prompt message, emits a sound or lowers the windows to alert the driver. The vehicle control priority can be temporarily adjusted to initiate braking and avoid risks [7]. In previous surveys, with the help of such precise driver distraction detection systems, the likelihood of vehicle accidents on the road can be reduced by 10% to 20% [8].

The NHTSA defines driver distraction as “any activity that diverts a driver's attention away from the primary task of driving.” The Centers for Disease Control and Prevention (CDC) also defines distracted driving more broadly [9]. When a driver's attention is diverted from the driving task to another activity, their attention is considered to be distracted. Cognitive distraction, behavioral distraction, and visual distraction are three typical types of driver distraction [10]. The driver is the primary decision maker and controller of the vehicle [11], so behavior distractions can greatly impact normal driving and easily result in traffic accidents. In recent years, the development of in-vehicle information system (IVIS) has made

the infotainment and communication functions in vehicles smoother and visually rich, further attracting the driver's attention [12]. On the other hand, daily life also requires people to have a great need for communication with their phones, and some drivers even need to frequently answer calls and send texts while driving [13]. These factors have also caused the driver's focus to become more and more dispersed.

Among the methods for driver distraction detection, researchers commonly use methods, such as physiological signals [14], [15], [16], vehicle information [17], [18], [19], and computer vision [20], [21], [22]. Most physiological signal detection methods have high-hardware costs, and invasive physiological sensing sensors often affect the driving experience of drivers [23]. Methods based on vehicle information also have certain limitations. Hardware failures of vehicles and sensors, as well as external interference, can result in inaccurate information collection, affecting the accuracy of the system [24]. These methods also require extensive data processing and advanced algorithm support [25]. To address these issues, researchers have gradually shifted their focus to computer vision methods [26]. Computer vision-based techniques for detecting driver distraction are nonintrusive, thus they do not affect the normal driving of drivers, and they are not as inaccurate as vehicle information methods due to information errors [27]. However, most researchers primarily adopt supervised learning methods and traditional convolutional neural network (CNN) models. By surveying the research work in the field of driver distraction behavior detection, three existing challenges can be inferred.

- 1) The strong dependence on labeled data sets for driver distraction behavior detection based on supervised learning is a major drawback. Supervised learning models require a large amount of labeled data for training, which significantly increases the cost of training. Additionally, labeling the data set requires a significant amount of manpower and resources. Furthermore, the complexity of real-world driving scenarios makes it difficult to accurately label data, increasing the difficulty and cost of creating a labeled distraction driving data set.
- 2) The traditional CNN-based driver distraction behavior detection model is inefficient in feature extraction and has limited ability to capture overall image information. CNN models only stack convolution and pooling operations. As a result, it is inefficient in capturing and representing long-term dependencies in the data. Additionally, CNN tend to focus on low-level image features, while the task of detecting driver behavior requires better control of global information. In images of driver distractions, basic visual pixels are correlated into objects, and the spatial relationships between objects form scene information. In such a visual detection task that requires behavior judgment, mastering high-level visual semantic information is more important.
- 3) Supervised learning models based on CNN have limited capability to learn representations and show an average generalization ability. Moreover, they are not effective at focusing attention on crucial regions to detect distractions. Furthermore, their dependence on labeled data

sets decreases their ability to generalize and transfer to novel tasks. This makes it challenging to ensure accurate recognition in more intricate and diverse driving scenarios.

Due to the effectiveness of prediction-based self-supervised learning in computer vision tasks, this article explores the application of the masked image modeling (MIM) self-supervised learning framework in the field of driver distraction behavior detection. Instead of using traditional supervised learning for model training and learning, we proposed a combined approach of pretraining and transfer learning. This approach effectively addressed the low generalization and substantial human consumption of labeled data in existing supervised learning methods. Instead of using CNNs, the Swin Transformer [28] is used as the encoder in the framework, taking advantage of the MSA mechanism to overcome the insufficient global feature-capturing ability of existing CNN models. A large and high-dimensional classification data set is used for pretraining, which provides a strong foundation for model representation learning and also makes the transferred model more robust and capable of better generalization. This will lead to better adaptability for recognition in various driving scenarios.

The main contributions of our work are summarized as follows.

- 1) In order to address the high-research cost and consumption issues associated with current supervised learning methods, we introduce a self-supervised learning method based on MIM for the detection of driver distraction behavior, named SL-DDBD. This will alleviate the time and effort consumption of data labeling. Meanwhile, through transfer learning, the model becomes more adaptable in downstream task detection scenarios, compared to the original supervised model, it has more effective feature attention and strong generalization ability.
- 2) In this work, we employ the high-performing Swin Transformer as the encoder for the self-supervised learning framework. In the transfer learning process, we consider the depth of the Swin Transformer structure and its relationship with the application scenario. Taking into account the features of the driver distraction detection task, the number of Swin Transformer blocks in stage 3 of the encoder structure is reduced to effectively decrease the computational cost and parameters of the massive encoder structure. At the same time, in order to balance the computation of the MSA [29] mechanism module and the improvement of the detection performance, the number of heads for W-MSA and SW-MSA [28] detection in each stage is redistributed. This adjustment alleviates the computational redundancy of the encoder in more detail while providing better recognition for driver distraction behavior detection.
- 3) In the model training, the best masking strategy was applied to improve the detection performance, and the proposed data augmentation strategies in this article were used for learning and training. Through experiments, a masking strategy with patch size of 64 and ratio of 0.5 was selected, which resulted in a more accurate

recognition performance. A multistrategy data augmentation approach was introduced in this work, including color jitter, motion blur, Gaussian noise, Mixup [30], and Cutmix [31], to better simulate a more diverse and realistic driving scenario, and improve the diversity of images. This enhances the robustness and generalization ability of the model.

- 4) This article compares the performance of self-supervised learning models and supervised learning models on the same data set and training configuration through visual comparison. The results further validate that self-supervised models are better at capturing high-level semantic information and have more focused feature attention on critical discrimination regions. The work also compares its results with the state-of-the-art supervised and unsupervised learning methods in driver distraction behavior detection, demonstrating the advanced and feasibility of the proposed method.

II. RELATED WORK

In recent years, there has been significant progress in related research fields, such as CNN [21], [32], [33], [34], attention mechanism [35], [36], [37], knowledge distillation [38], [39], [40], and Transformer [41]. Depending on the advances in these related fields, researchers have combined these methods with research areas. This has led to a wide range of developments in driver distraction behavior detection.

A. Driver Distraction Detection

Driver distraction detection is a hot topic of ADAS. Many traditional methods detect driver distraction by integrating CNN or graph neural network (GNN) models and have achieved good results. Hu et al. [32] proposed MVCNet, which processes information from three visual contents, including texture features, and incorporates GNN and CNN, optical flow, and semantic segmentation information. They also created a new semantic attention module that is integrated into both the CNN and graph attention network (GAT) branches. The multibranch model decodes the integrated features to finally generate the driver's gaze attention map. Different from previous work, Xing et al. [33] designed a system that can simultaneously recognize a driver's body activities and mental state. The system is based on a framework of deep encoders and decoders. The encoder is designed based on CNN to extract effective spatial information from driving activity videos, while the encoder is designed using a fully connected network and LSTM-based RNN for estimating different driving states.

The areas of computer vision have extensively used attention mechanisms, which have improved the overall performance of networks and greatly helped with feature extraction and classification. Huang et al. designed a deep 3-D network (DhRN-AMED) for driver distraction detection. They inserted the attention mechanism as a nonlinear transform layer into the residual network with a soft threshold to extract features related to driver behavior. Li et al. [36] used an object detection method to implement recognition of driver behavior and

proposed an AB-DLM network model for driver distraction detection. They stacked SE attention mechanism module in the architecture and employed bi-directional feature pyramid networks (BiFPNs) instead of the original path aggregation network (PANet) as a new multiscale fusion network.

On the other hand, the lightweight of driver distraction detection model is very important for deployment to mobile devices. Some researchers have shifted their focus to knowledge distillation. Liu et al. [38] created a high-performance teacher network. Knowledge distillation was then used to direct the student network's learning process after gradually enhancing CNN's robustness to illumination shifts from shallow to deep layers. The distilled knowledge is transferred from the teacher network to the student network, resulting in a student network with 2.03M parameters. This is a powerful method for lightweight the driver behavior recognition model. Transformer is also gradually growing in the computer vision field. However, most researchers have not yet done much work in driver behavior detection. Wang et al. [41] proposed a detection model that integrates both CNN and Transformer architecture, incorporating specific enhancements to the Transformer structure for better performance. The original multilayer perceptron (MLP) module is replaced with a convolution module to reduce the number of model parameters and improve detection speed. A label-smoothed loss function is designed and applied to the model learning.

B. Self-Supervised Learning for Masked Image Modeling

Initial study and experimentation on self-supervised learning was conducted in the area of natural language processing before gradually expanding to computer vision. [42]. Nowadays, self-supervised learning using MIM has slowly gained attention [43], [44], [45], [46], [47].

Bidirectional encoder representation from image Transformers (BEiT) [43] is a pioneering work that transfers the BERT-style pretraining method to the visual domain and introduces the concept of MIM pretraining, making important contributions to the field of self-supervised learning. BEiT first annotates the original image and randomly masks image blocks, feeding the masked image into the encoder, with the main pretraining goal being to recover the masked image blocks based on the unmasked ones. Masked autoencoders (MAEs) proposed by He et al. [44] is a highly influential work that drives the development of MIM self-supervised learning. MAE compared to BEiT, advocates a simpler training logic. It first proposed random masking of images and directly reconstructed masked image blocks for training. Adopted an asymmetric structure of encoder-decoder, where the encoder only calculated nonmasked image blocks and used a lightweight decoder design. Soon after MAE was proposed, Xie et al. [45] proposed a simpler mask learning framework. Regressing to a more original and simple design, including random masking, regressing RGB pixel values, and directly using a single linear prediction head. A simple framework for MIM (SimMIM) also achieved good results, achieving a top accuracy of 87.1 on ImageNet-1k. Later, Chen et al. [46] noticed some shortcomings in previous

work, such as a focus on semantic learning in image center blocks and neglect of edge regions, which also made context autoencoder (CAE)'s contribution one of the models not only focus on the center area of the image. Second, CAE splits representation learning and front-end tasks, making the encoder more focused on representation learning.

The UDL model proposed by Li et al. [48] used a traditional contrastive unsupervised learning method. In the UDL model, a new backbone and projection head were constructed using MLP. They also designed and used a loss function with a stop-gradient strategy to guide learning and training, resulting in a more robust model. Although this was a good attempt, research on self-supervised learning with the MIM method in the field of driver behavior detection is still scarce.

III. METHODS

A. Suitable Self-Supervised Learning Framework

Subsequently, we propose a self-supervised learning method based on MIM for detecting driver distraction behavior. The pipeline of the proposed SL-DDBD is shown in Fig. 1. The proposed method consists of two stages: 1) pretraining and 2) fine-tuning. In the pretraining stage, a large unlabeled data set is used for self-supervised learning. The MIM method is employed to mask a part of the input image and train the model to repredict the masked area of the image. The fine-tuning stage uses transfer learning to fine-tune the model with a specific data set. This stage includes precision optimization and lightweight encoder, as well as data augmentation. The best masking strategy is determined through experiments. The fine-tuned model is capable of adapting to downstream tasks with better adaptability. The self-supervised learning pretraining framework consists of four main parts.

- 1) *Masking Strategy*: This section focuses on how to choose the masked part of the image and considers the overall level of masking and the size of the masked image block. The output is then passed to the next section.
- 2) *Encoder Architecture*: A strong encoder with good learning ability can be used for various visual downstream tasks. The encoder extracts the latent representation of the masked image and predicts the original signal of the masked part through training and learning.
- 3) *Prediction Head*: The prediction head is used for the latent representation to represent the original signal of the masked patch.
- 4) *Prediction Target*: This part specifies the form of the original signal to be predicted, which can be either the original pixel values or the transformation of the original pixels. The loss type is also defined in this part. Cross-entropy categorization loss and L1 or L2 regression loss are common choices.

In the masking strategy, a random masking strategy is used. Image blocks are the basic processing units in the entire system framework. Using this as a basic unit can facilitate masking operations at the block level. It is convenient and easy to implement masking, complete masking, or no masking on the customizability of the image. For different encoders, there are different patch sizes. For the Swin Transformer encoder, we

consider patch size dimensions at different resolution levels, ranging from 4×4 to 32×32 . For the vision Transformer (ViT) encoder, a default patch size of 32×32 is used.

In the prediction head, it is necessary to ensure that the input to the prediction head is consistent with the output of the encoder. Once the prerequisites are met, the form and size of the prediction head can be customized. The prediction head is then defined to predict the target. In some early works, an autoencoder was followed with a heavy prediction head. The use of a complex detection head did not result in better performance but rather increased the training cost. In this paper, considering the application scenario of driver distraction detection, a simple prediction head is used to accomplish this task. We use 1×1 convolution kernels to implement a single linear layer to predict pixel values.

In the prediction target, the pixel values are continuous in the color space, and the original pixels of the masked area are predicted directly through regression. Each feature vector in the feature map is mapped back to the original resolution and is in charge of forecasting the original pixels. This is used to predict all the pixel values of the input image at all resolutions.

We apply a 1×1 linear layer with an output dimension of $3072 = 32 \times 32 \times 3$ to the 32×32 down-sampling feature map made by the Swin Transformer encoder to depict the RGB values of the 32×32 pixels. In order to account for lower resolution objects, the original image is down-sampled at multiple dimensions, respectively. These include $\{32 \times, 16 \times, 8 \times, 4 \times, 2 \times\}$.

Use ℓ_1 -loss on the masking pixels

$$L = \frac{1}{\Omega(x_M)} \|y_M - x_M\|_1 \quad (1)$$

where, $y \in \mathbb{R}^{3HW \times 1}$ are the input RGB and predicted values, M is the collection of masked pixels, and $\Omega(\cdot)$ is the number of elements.

B. Encoder Architecture

Due to the high requirements of the encoder to extract the latent feature representation of the masked part of the image, the encoder needs strong representation learning ability. Therefore, we consider using the Swin Transformer as the encoder in the self-supervised learning framework.

Transformer has shone in the field of computer vision after being transferred from the NLP field, gradually becoming the main general pillar of computer vision. However, After ViT, the birth of Swin Transformer was a milestone that pushed the development of Transformer for computer vision tasks [49]. Swin Transformer is more suitable for visual tasks, with two important design points: 1) hierarchical Transformer and 2) shifted window, resolve large scale differences in visual entities and the high-pixel resolution of images compared to text.

The architecture of encoder as shown in Fig. 2. First, the image is input into the patch partition module for block processing. The image is divided into multiple patches of 4×4 pixel size. Then, it is flattened in the channel direction. When an RGB three-channel picture is used as the input, then each patch has $4 \times 4 = 16$ pixels, and each pixel has R , G , and B

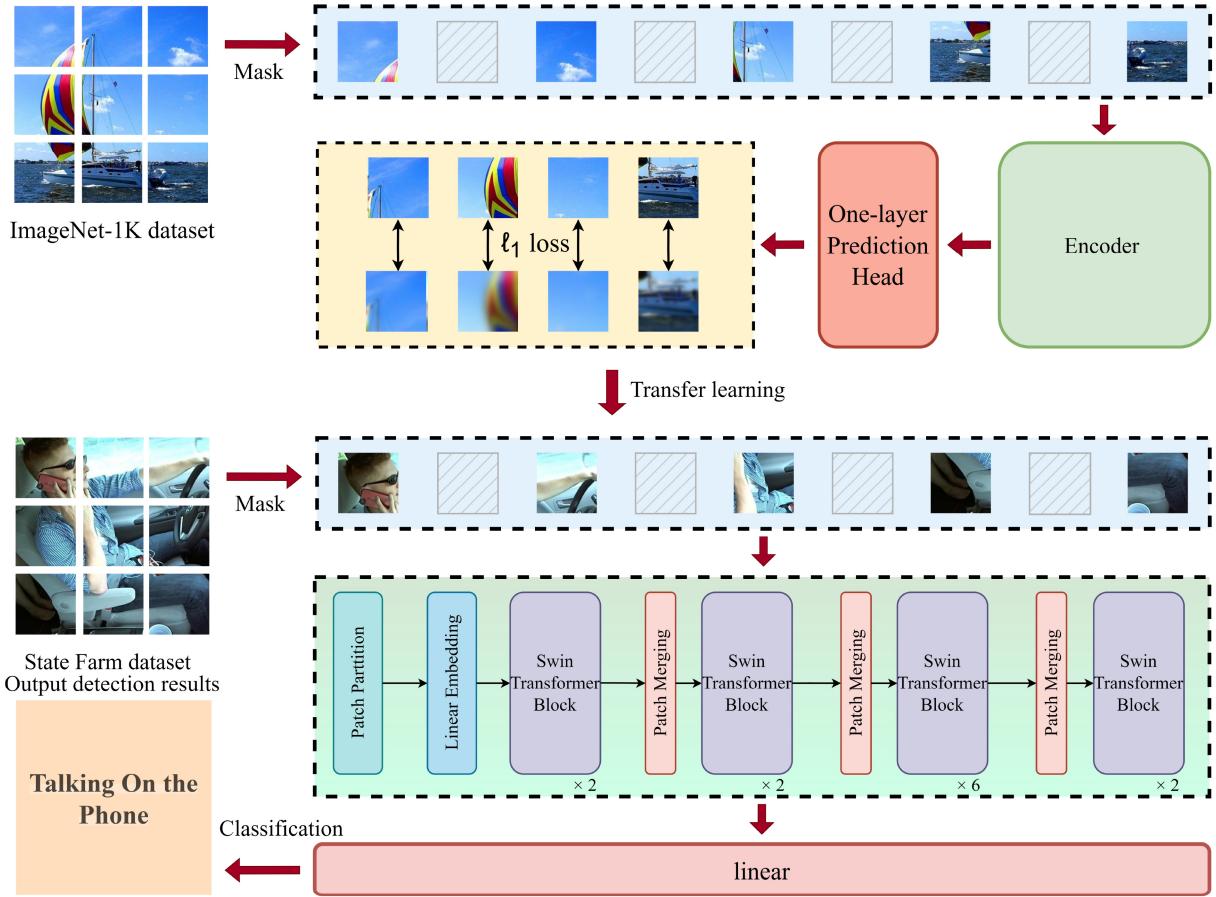


Fig. 1. Pipeline of the proposed SL-DDBD.

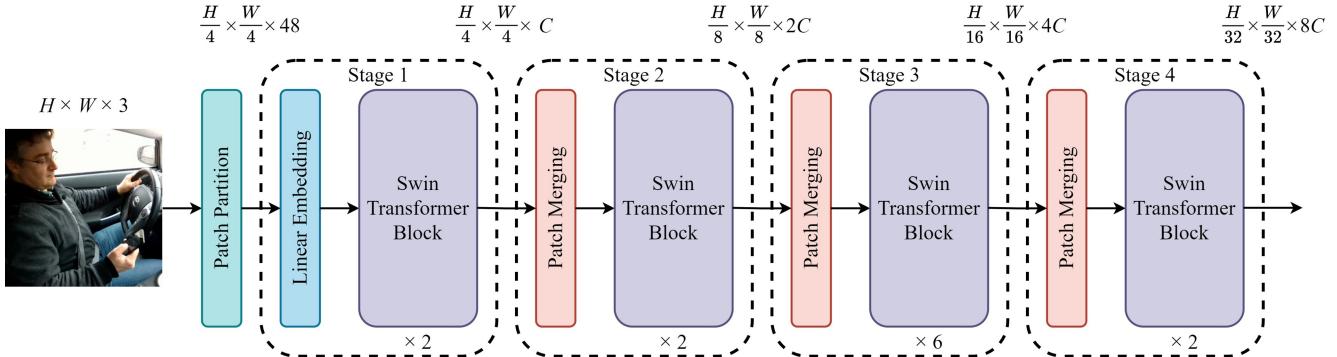


Fig. 2. Architecture of encoder in the self-supervised learning framework.

three values, so after flattening, it is $16 \times 3 = 48$. After processing by the patch partition, the image shape changes from $[H, W, 3]$ to $[(H/4), (W/4), 48]$. The linear embedding layer linearly transforms each pixel channel, that is, from the original 48 to C . C value is defined. After this layer, the image shape is $[(H/4), (W/4), C]$.

The image is then processed through four stages one by one and the size changes accordingly. Linear embedding layer is only in stage 1. Patch merging layers and various numbers of Swin Transformer blocks make up the final stages. The Swin Transformer block, as the main working structure, contains two structures. One is the block using the W-MSA module,

and the other is the block using the SW-MSA module. So when stacking the Swin Transformer blocks, they are stacked in pairs.

The patch merging process is shown in Fig. 3. When the input is a single-channel feature map of 4×4 size, patch merging divides each 2×2 neighboring pixel into a patch. By connecting the pixels in the same location in each patch, four feature images are created. Concatenating four feature maps in the depth direction, followed by passing through a LayerNorm layer. Finally, a fully connected layer performs linear transformation on the depth direction of the feature map, reducing the depth of the feature map from C to $(C/2)$. As a

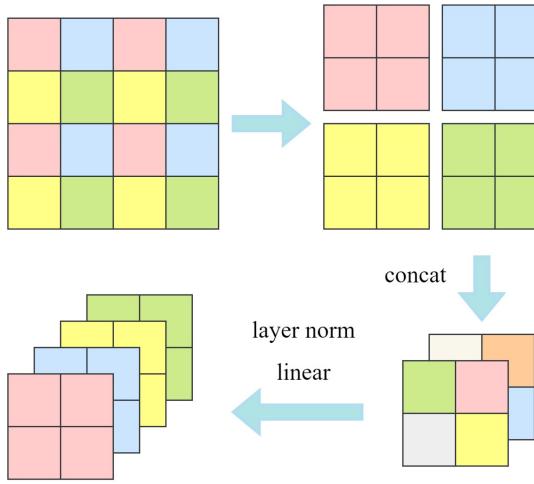


Fig. 3. Workflow of patch merging.

result, after stages 2, 3, and 4, the shape of the image will be halved in both width and height while the number of channels will be doubled.

C. Improved Data Augmentation Strategy

In terms of the data set, it is dangerous to create driving scenarios that simulate actual driver distraction. Therefore, the number of real driver distraction data sets is relatively small, and the quality of the data sets still needs to be improved. The quantity and diversity of the data set for deep learning network models often greatly affect the training results of the model, including the recognition accuracy and robustness. To tackle this problem, this work presents a data enhancement strategy that aims to simulate and closely mimic real-world driving scenarios to increase the quantity and diversity of the data set.

- 1) *Color Jitter*: An effective data augmentation method, randomly changing the exposure, saturation, and hue of the images. The aim is to simulate the driving situations of the driver in the driving room under different lights and weather conditions. This increases the number and diversity of training data sets, allowing the model to learn the differences brought by changes in lighting.
- 2) *Motion Blur*: Turns a clearly focused image into a motion blur effect. Most of the images are captured in focus for easier feature extraction. In realistic conditions, the driver's movements are fast and continuous in the video of the detection device. Therefore, it is necessary to consider the phenomenon of motion blur caused by the driver continuously performing actions.
- 3) *Gaussian Noise*: A form of noise whose probability density function has a Gaussian or normal distribution is referred to as Gaussian noise. It applies overall noise to images in order to enhance their diversity and improve the model's ability to learn representations.
- 4) *Horizontal Flipping and Random Scaling*: Random scaling enhances the model's ability to detect distracted driving actions of multiple scales, while horizontal flipping better simulates the driving scenarios of right-side and left-side vehicles.

- 5) *Cutmix*: A random crop box is generated, and a corresponding portion of the A image is cropped. Then the corresponding ROI in the B image is placed in the cropping area of the A image to form a new image. The loss is also solved by weighted summing [31]. Employing hard fusion of two images and concurrently implementing soft fusion strategies for labels, Cutmix ensures the distribution of the data set remains unaltered. Let $x \in \mathbb{R}^{W \times H \times C}$ and y represent the training images and their labels. The merge operation is defined as

$$\tilde{x} = \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B \quad (2)$$

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \quad (3)$$

where $\mathbf{M} \in \{0, 1\}^{W \times H}$ represents a binary mask indicating the deleted and filled positions in the image. $\mathbf{1}$ is a binary mask filled with 1, and \odot represents elementwise multiplication. λ belongs to the Beta(α, α) distribution, and if $\alpha = 1$ is set in the experiment, λ follows a uniform distribution between (0, 1). To sample the binary mask \mathbf{M} , it is necessary to sample the bounding box $\mathbf{B} = (r_x, r_y, r_w, r_h)$ of the cropping area. Then, sample x_A and x_B based on the sampling result for cropping and padding. In the experiment, the aspect ratio of the rectangular mask \mathbf{M} is proportional to that of the original image. The bounding box coordinates are sampled as follows:

$$r_x \sim \text{Unif}(0, W), r_w = W\sqrt{1 - \lambda} \quad (4)$$

$$r_y \sim \text{Unif}(0, H), r_h = H\sqrt{1 - \lambda}. \quad (5)$$

The cropping area ratio is $(r_w r_h / WH) = 1 - \lambda$, $\mathbf{M} = 0$ in the cropping area \mathbf{B} , while the rest of the area has $\mathbf{M} = 1$.

- 6) *Mixup*: A regularization technique that randomly blends the pixels of two training images to create a new image that incorporates the labels of both input images [50]. This approach is designed to enhance the diversity and complexity of the training data set by maximizing the combination of different contextual information, thereby improving performance. The calculation formula is as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \quad (6)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \quad (7)$$

where $\lambda \in [0, 1]$ follow a Beta(α, α) distribution, α is a hyperparameter that controls the interpolation strength. The larger the value of α , the more obvious the interpolation effect. While the smaller the value, the closer the Mixup enhancement effect tends to be ineffective.

D. Self-Supervised Pretraining

In self-supervised learning, the training process is similar to a conventional autoencoder. In this work, the input information is mapped to latent features representation using a Swin Transformer encoder. Then the liner layer is used as a prediction head to reconstruct the masked part of the image by potential feature information. During pretraining, large-scale,

unlabeled data sets are used for representation training and learning. In this work used Imagenet-1K data set, which contains no labels. The images in the data set are subjected to a random masking strategy in the self-supervised learning framework, with patch size set to 32 and ratio set to 0.5. The inputs are fed into the encoder, which processes both the visible tokens and the masked tokens.

The encoder's core working module is the Swin Transformer block. In the Transformer, the self-attention module is the basic operational unit, just like the convolutional operation in a CNN network. The self-attention mechanism in the Transformer block is able to adaptively model the long-term dependency relationships among sequence markers. One type of dependency relationship can only be established by a particular attention function. MSA aims to learn multiple dependencies from different representation subspaces. In particular, the keys, values and queries of the d_{model} dimension are divided linearly into h groups, the attention function is carried out in all groups concurrently, generate d_{model}/h dimension output values and these values are concatenated and projected.

MSA is expressed as follows:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(y_1, \dots, y_n)W^O \quad (8)$$

$$y_i = \text{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}}\right) \mathbf{V}_i \quad (9)$$

where $d_k = d_{\text{model}}/h$ is the average feature area of each attention head. $\mathbf{Q}_i = \mathbf{X}_i \mathbf{W}_i^q$, $\mathbf{K}_i = \mathbf{X}_i \mathbf{W}_i^k$, and $\mathbf{V}_i = \mathbf{W}_i^v$ stand for the key, query and value. W^O is the learnable projection matrix, and h is the total number of self-attention head. \mathbf{W}_i^q , \mathbf{W}_i^k , and \mathbf{W}_i^v are parameter matrices, and \mathbf{X}_i represents the feature matrix of the i th head.

E. Transfer Learning With Accuracy Optimization and Light Weighting

In the field of computer vision, large networks are designed to provide better service for high-difficulty image classification tasks and object detection. However, these deep and wide networks also drive the progress of various visual downstream tasks. Transfer learning is an important part of realizing downstream tasks [44]. We consider the relationship between complex networks and the specific application scenario in this work. In order to achieve better detection results, we also need to consider the deployment of the model to hardware and specific engineering applications. Therefore, we perform lightweight and precision optimization tuning work on the encoder in the self-supervised learning framework.

The Swin Transformer block is shown in Fig. 4. The image input enters the first block, where the image needs to go through a layer norm layer and a W-MSA module, with a skip connection alongside both processing steps. The image continues to enter the layer norm layer and MLP module, with a skip connection alongside this path as well. At this point, the image has completed processing through the first block and outputs to the second block. The second block is similar in overall structure to the first, but instead of using the conventional W-MSA, it uses SW-MSA. The consecutive Swin Transformer block calculation process is as follows:

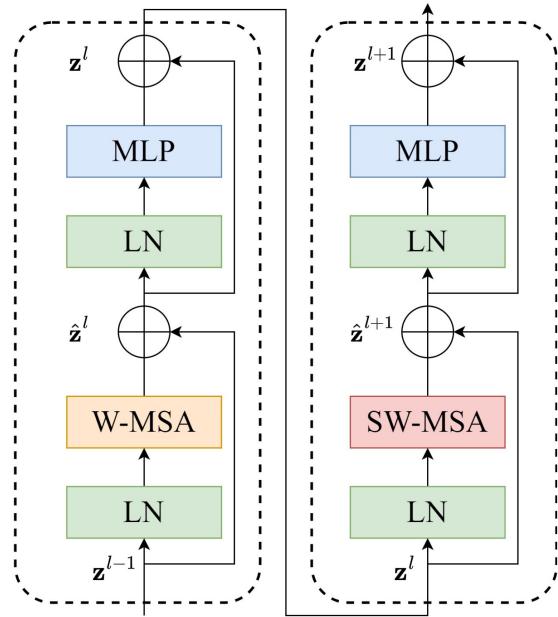


Fig. 4. Network of Swin Transformer block.

$$\widehat{\mathbf{z}}^l = W - \text{MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1} \quad (10)$$

$$\widehat{\mathbf{z}}^l = \text{MLP}\left(\text{LN}\left(\widehat{\mathbf{z}}^l\right)\right) + \widehat{\mathbf{z}}^l \quad (11)$$

$$\widehat{\mathbf{z}}^{l+1} = SW - \text{MSA}\left(\text{LN}\left(\mathbf{z}^l\right)\right) + \mathbf{z}^l \quad (12)$$

$$\widehat{\mathbf{z}}^{l+1} = \text{MLP}\left(\text{LN}\left(\widehat{\mathbf{z}}^{l+1}\right)\right) + \widehat{\mathbf{z}}^{l+1} \quad (13)$$

where $\widehat{\mathbf{z}}^l$ and $\widehat{\mathbf{z}}^{l+1}$ denote the output features of the (S)W-MSA module and MLP module of block l , respectively; W-MSA and SW-MSA represent the multiheaded attention from the partitioned configuration with regular window and shifted window.

In the encoder, each image goes through W-MSA and SW-MSA processing in Swin Transformer block. Therefore, the number of Swin Transformer blocks must be even. Compared to large and difficult visual tasks, such as world object classification, driver distraction behavior classification is less complex and has more distinctive features. Although large networks perform well in recognition of complex classification tasks, overly complex networks are not the best solution for relatively simple detection scenarios. In order to reduce the computation and parameters of the massive encoder, the work changes the number of stacked Swin Transformer blocks in stages of the encoder structure. Specifically, the original 18 stacked Swin Transformer blocks in the third stage are reduced to 6.

The W-MSA and SW-MSA are the main working modules. Compared to the original MSA, the Swin Transformer block performs MSA on the window. The MSA, due to its window-based setting, significantly reduces the computation cost compared to the original. The computation cost formula of MSA is as follows:

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C \quad (14)$$

where Ω is the computation, h and w are the height and width of the image respectively, and C is the number of channels. W-MSA module divides the feature map into a

TABLE I
DRIVER BEHAVIORS IN THE STATE-FARM DATA SET

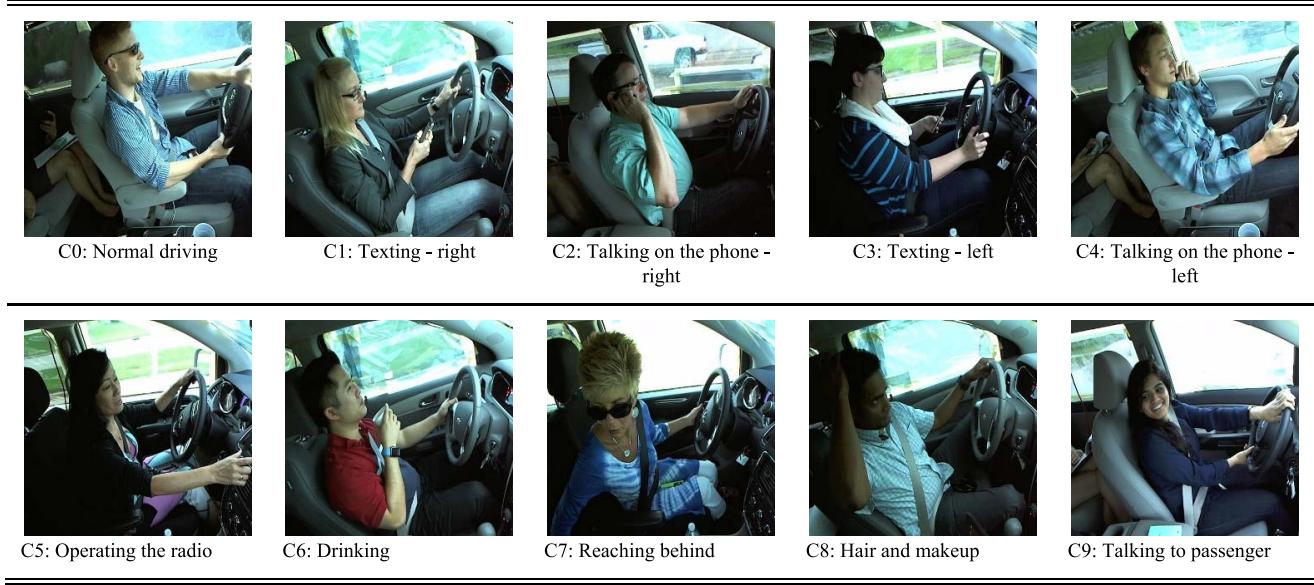


TABLE II
DATA SET DIVISION

	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
(a)	Train	1991	1814	1854	1877	1861	1850	1860	1602	1529
	Test	498	453	463	469	465	462	465	400	382
	Total	2489	2267	2317	2346	2326	2312	2325	2002	1911
(b)	Train	9156	8268	8468	8584	8504	8448	8500	7208	6844
	Test	2000	2000	2000	2000	2000	2000	2000	2000	2000
	Total	11156	10268	10468	10584	10504	10448	10500	9208	8844

window with the width and height of M . A feature map that will get $(h/M) \times (w/M)$ windows, and then use the multiheaded self-attention module for each window. Since the window's width and height are M , bring the above formula as $4(MC)^2 + 2(M)^4C$, the final W-MSA calculation is

$$\Omega(W - \text{MSA}) = 4hwC^2 + 2M^2hwC. \quad (15)$$

It is similar that the number of heads in the MSA is controllable and the size of the encoder is positively correlated with the number of heads. The number of heads in the W-MSA and SW-MSA in the four stages of the encoder are set differently. The original baseline sets the number of heads to $\{4, 8, 16, 32\}$. However, in the case of very few categories, the redundant number of heads brings more computational parameters and does little to enhance the detection task for fewer categories. Hence, we adjust the number of detection heads in each stage to $\{3, 6, 12, 24\}$.

IV. EXPERIMENTS

A. Data Sets and Comparison of Each Data Augmentation Strategy

The data set used in this work is the State Farm data set from the official Kaggle competition [51]. This data set is a comprehensive and diverse data set for driver behavior monitoring, which includes 26 participants of different races, skin colors, and genders (13 male and 13 female) from America,

Asia, and Africa. All images in the data set were captured by a camera fixed in the car dashboard, and all images are RGB pixels. The data set consists of a total of 22 424 images. As our initial data set, we randomly divided each class of images into 80% for training and 20% for testing.

We used the proposed data augmentation strategy to enhance the data set to improve the generalization ability and robustness of the proposed detection algorithm model. After data augmentation, the training images are 81 976 and the test images are 20 000. The specific classification of distracted driver behavior in the State-Farm data set is shown in Table I. The basic data set division is shown in Table II(a), and the division of the expanded data set using our proposed data augmentation strategy is shown in Table II(b).

In order to explore the positive effects of the data augmentation strategy proposed in this work on driver distraction detection, we evaluated the effectiveness of each data augmentation strategy. This exploration aims to discover the potential connection between different data augmentations and driver distraction detection. Under the same software and hardware environment, we set the unified masking strategy of patch size 64, ratio 0.5, and used the improved encoder. We trained seven models with each of the six data augmentation strategies and without any data augmentation. The performance of the seven models was tested on the same data set and the results are shown in Fig. 5.

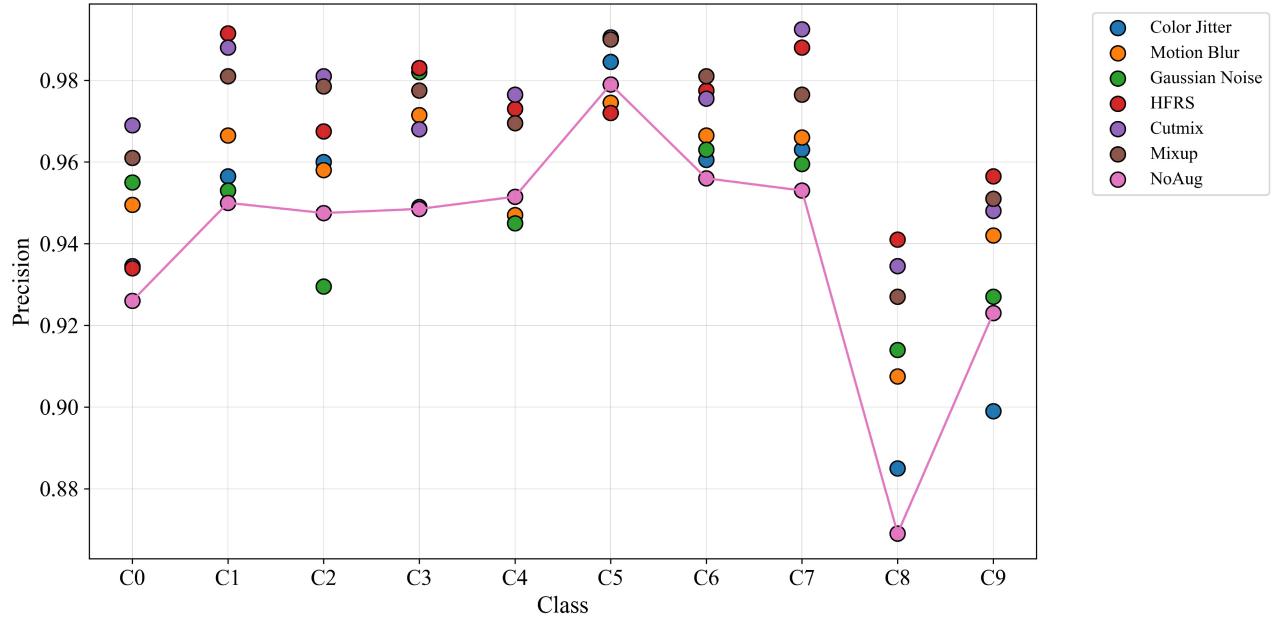


Fig. 5. Comparison of precision of models using different data enhancement strategies.

TABLE III
HARDWARE AND SOFTWARE CONFIGURATION

Operating systems	Linux Ubuntu 20.04.2 LTS
CPU	Intel Xeon E5-2680 v4
GPU	NVIDIA Geforce RTX 2080 ti (12G)
Memory	16G × 2
Solid state drives	480G
Pytorch	V1.10
CUDA	V10.0

The NoAug model (without any data augmentation strategy) has the worst overall performance, with the lowest precision in multiple categories, including C0, C1, C3, C6, C7, and C8. The models with Cutmix and Mixup perform the best among all data augmentation strategies, followed by horizontal flipping and random scaling (HFRS) and motion blur, and finally Gaussian noise and color jitter. The precision of the models improved significantly after data augmentation. For example, the NoAug model has the lowest recognition precision in category C8, but the precision in this category for all other models augmented has improved. The precision of the motion blur and Gaussian noise models is slightly degraded on C2 and C4, both of which belong to the “talking on the phone” category. Motion blur and Gaussian noise affect the images that have few phone pixels in the image, so there is a small decrease in precision in both categories. However, the overall accuracy is still higher than that of NoAug model. Since the Cutmix and Mixup data augmentation methods remove and overlap some parts of the image, causing significant changes in the image. These data augmentation strategies increase the training difficulty but also improve the model’s feature extraction from the image, resulting in highly robust and accurate models after training. Next in line for effectiveness are the HFRS and motion blur. Like the previous data augmentation strategies, HFRS also makes significant changes to the image by randomly altering its size and scale, thus enhancing effective

feature extraction. The improvements in precision brought by Gaussian noise and color jitter are not obvious because they only induce simple changes to the image by adding noise and color variations. The proposed data augmentation strategy in this work has a positive effect on the detection of distracted driver behavior. It enhances the diversity of the data set, achieves data expansion, and improves model detection precision.

B. Training Details

This article uses the experimental environment configuration in Table III to guarantee the effectiveness of model training and testing. A large-scale self-supervised pretraining is conducted using the ImageNet-1k data set. During pretraining, the size of input images is adjusted to 192×192 , and the window size is adjusted to 6 to adapt to the changed input image size. In self-supervised pretraining, the training cycle is 800 epochs, and AdamW optimizer with cosine learning rate adjustment is used. The specific training hyperparameters are as follows: batch size is 2048, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay is 0.05, and the base learning rate is 8e-4. In the masking strategy for MIM: the random masking strategy is used, the masked ratio is 0.5, and the patch size is 32. The predicted target image size in the linear prediction head is 192×192 .

In the transfer learning process, we used both the basic version and the augmented version of the Sate-Farm data set. In the hardware and software environment, we used six Intel Xeon E5-2680 v4 CPUs and an NVIDIA RTX 2080 TI GPU with a 10-G Memory size. We also used the AdamW optimizer for transfer learning, with all training periods changed to 110 epochs, the batch size adjusted to 32, the base learning rate changed to 5e-3, and the weight decay remaining at 0.05. In the masking strategy for image modeling, we used a random masking strategy. The masked ratio and patch size were used as variable parameters in the experiment. The image input size

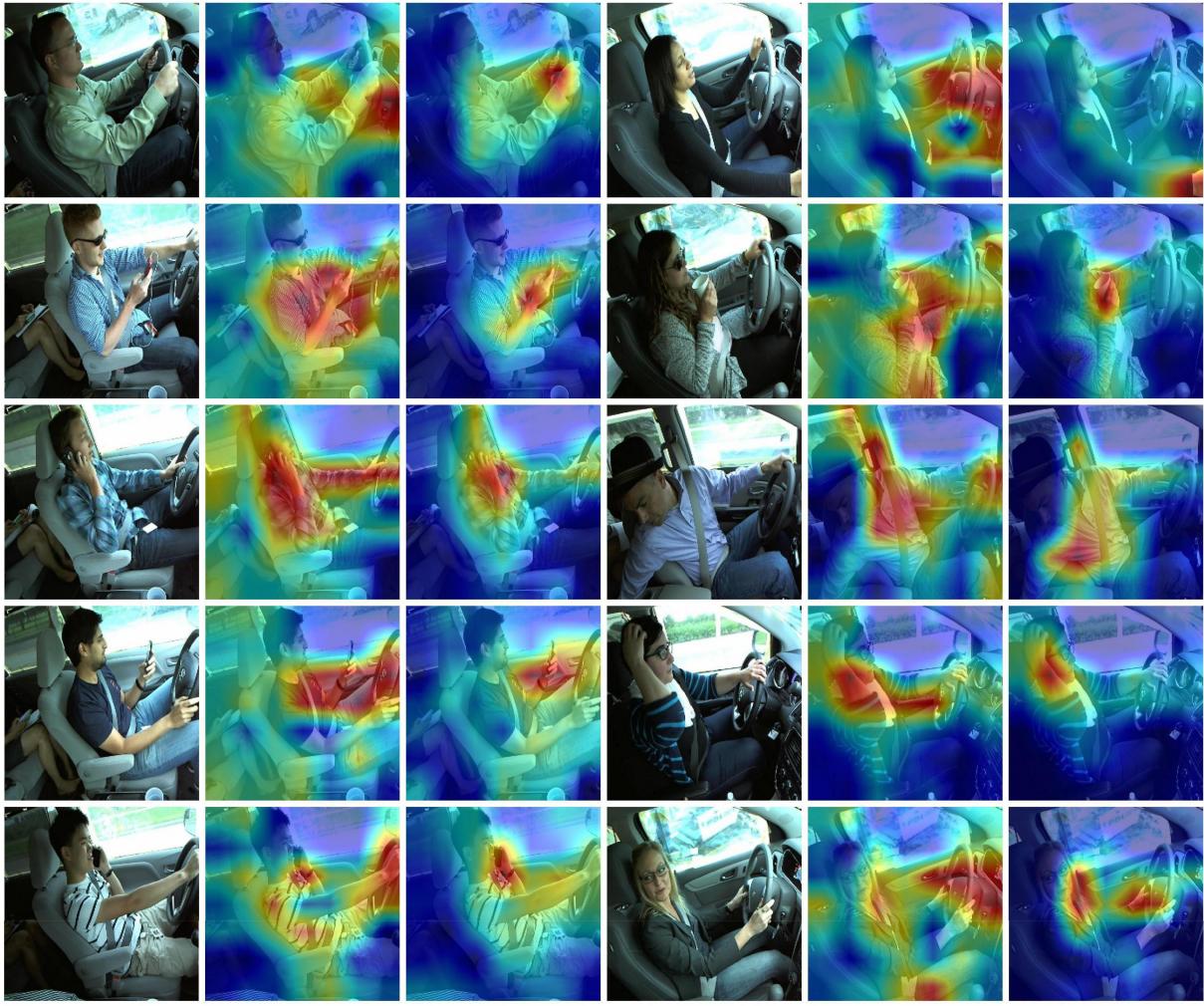


Fig. 6. Self-supervised and fully supervised learning visualizations on state farm data set.

was uniformly changed to 224×224 , and the window size was adjusted to 7.

C. Self-Supervised Learning and Masked Image Modeling Visualization

To investigate the advancement of self-supervised learning based on MIM, we visualize the self-supervised learning model and the supervised learning model, and the results are shown in Fig. 6. We adopt the same encoder structure, training settings, and data set. We trained a supervised learning model. The left column shows the original images of the driver distraction behavior data set, the middle column shows the visualization effects obtained by the supervised model, and the right column shows the visualization effects obtained by the self-supervised model in this work. We use the grad-cam method [52] for visualization and use gradients for visualization. We select the LayerNorm layer of the last Swin Transformer block module in backbone.

The visualization comparison results in the ten types of driver distraction show that the supervised learning model's attention is relatively dispersed in all categories. For example, using a cell phone for calling and texting, the supervised

model overall focuses on a large area. At the same time, it also sometimes needs to pay attention to the situation on the steering wheel to distinguish distracted behavior. However, the self-supervised learning model has a very small focus area and concentrates on the key determining parts of the driver's behavior. This is particularly evident in the category of the driver drinking, where the self-supervised learning model focuses only on the driver's hand holding the cup in front. The same is true for distracted behavior using a phone; where only the hand and the phone part need to be focused on to classify the behavior. Overall, the self-supervised model's attention is focused on the key parts of the scene objects and has a better grasp of the feature information, avoiding the problem of feature redundancy and excessive computational costs.

D. Comparison of Transfer Learning Optimization

This experiment investigates the impact of self-supervised masking strategies on detection performance by setting different masking details for the baseline. The patch size is the size of the masked image blocks in the masking strategy. After being masked, all patches are input into the encoder, so the patch size directly affects the encoder's receptive field.

TABLE IV
MODEL PERFORMANCE WITH DIFFERENT MASKING STRATEGY

Patch size	Masked ratio	Acc(%)	Training time
16	0.4	88.08	8:34:21
16	0.5	89.33	8:40:49
16	0.6	87.05	8:45:26
32	0.4	84.61	8:51:42
32	0.5	84.92	8:55:30
32	0.6	88.86	8:57:49
64	0.4	83.73	8:52:57
64	0.5	86.68	8:54:19
64	0.6	88.70	8:53:06

A larger patch size provides a larger receptive field but also contains more irrelevant features. Smaller patches reduce the receptive field and limit the communication between the overall image information. These situations affect the feature learning of the encoder and result in poor classification performance. The masked ratio is also an important part of the masking strategy. It affects the degree of masking of the entire input image. A smaller ratio may cause the encoder to learn weak representations and make it difficult to restore and predict the masked image. A larger ratio value leads to too much masking, making it difficult to train and resulting in a strong representation ability. Therefore, the ratio value will greatly impact the model's performance. The masking strategy is determined by the patch size and ratio value. In this experiment, we will set different ratio values and patch size values to verify the impact of the masking strategy on the model's detection performance.

In the experiment, set the patch size first, and then adjust the size of masked ratio value. The specific test results of different masking strategy models are shown in Table IV.

Acc represents the accuracy, and training time is the training time under the hardware conditions used. The baseline model's mask setting is patch size 32, ratio 0.5. It adopted the pretrained model setting and encoder structure. The baseline accuracy is 84.92%. When the patch size is 32, the baseline is 0.31% higher than the masked ratio set to 0.4, and the masked ratio set to 0.6 is the highest accuracy of 88.86% at this size. When the patch size is 64, the masked ratio set to 0.6 has the highest accuracy of 88.70%. When the patch size and ratio value are larger, an accuracy of 88.70% can be achieved. This means that in a situation where the image is heavily masked and the model training is relatively difficult, the model has good learning representation ability. Even when the patch size is 16 and the masked ratio is set to 0.4, good performance results have been obtained. This means that in the case of less image masking, the model's feature learning for driver distraction detection scenes is favorable, and favorable recognition results have been obtained. The accuracy reached 89.33% when the masked ratio was 0.5. This is the highest accuracy obtained from the masking strategy experiment before optimization, compared to the baseline, it increased by 4.41%. The training time of all models fluctuated between 8:34:21 to 8:57:49. The longest training time did not

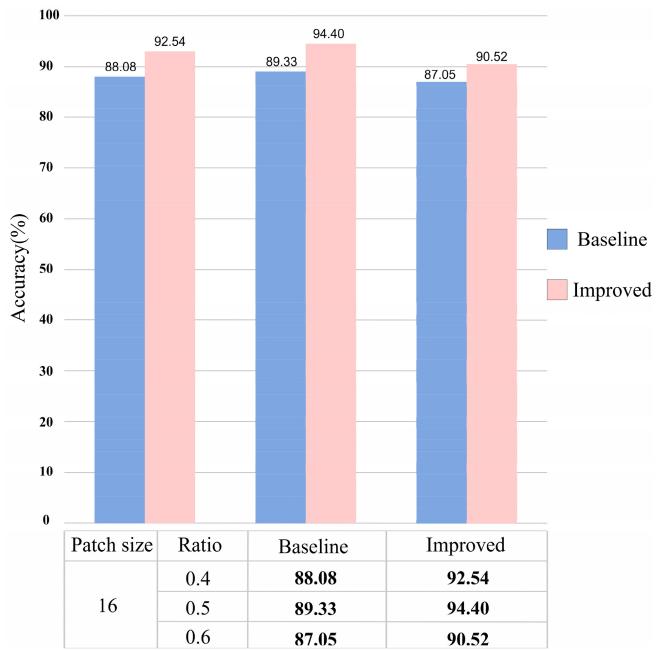


Fig. 7. Comparison of the accuracy of the baseline and improved models with a fixed patch size of 16 and different ratio values.

exceed 9 h. Therefore, overall, the change in patch size and masked ratio has a very small impact on the training cost of the entire model.

The experiment in this section will verify the impact of optimization in transfer learning on the model's detection performance. The optimized model will also experiment with different masking strategies. Figs. 7, 8, and 9 are the comparison results of the optimized model and the baseline in the case of patch size 16, 32, and 64, respectively. Each figure shows the accuracy of the two models under the same masking strategy, with the same patch size, presented in the form of a bar graph, making it more intuitive to show the accuracy improvement brought by the optimization work.

In Fig. 7, the patch size is 16 and all improved models increase accuracy by 4.46%, 5.07%, and 3.47% compared to the original model for each masked ratio. The original model has the lowest accuracy of 87.05% when masked ratio is

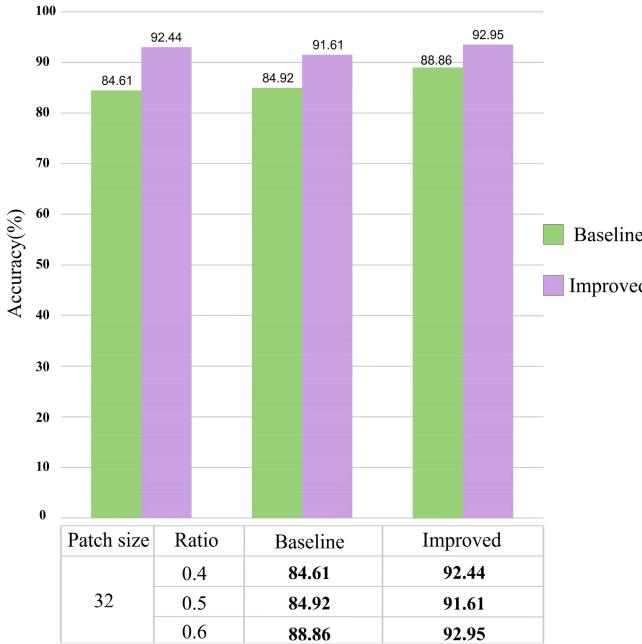


Fig. 8. Comparison of the accuracy of the baseline and improved models with a fixed patch size of 32 and different ratio values.

0.6. However, the optimized model still maintains an accuracy above 90% under this setting. At the same patch size setting, the optimized model reached an accuracy of 94.40% when masked ratio is 0.5, which is the highest accuracy under this setting.

The improved model in Fig. 8 shows an increase in accuracy of 7.83%, 6.69%, and 4.09%, respectively, compared to the original model. The baseline model has an accuracy of only over 84% when the masked ratio is 0.4 and 0.5. However, after the model is optimized, the accuracy improvement is significant, all reaching above 91%. Fig. 9 also shows that the performance of the improved model is still better than the original model. Especially when the masked ratio is 0.5, the accuracy of the improved model reaches 95.13%, which is the highest recognition accuracy among all masking strategies. Compared to the original model under the same strategy, it has increased by 8.45%, and compared to the baseline, it has increased by 10.21%. In all the above model comparisons, the performance of all improved models is better than the original model. The separate validation on various masking strategies indicates that the optimization work on accuracy in transfer learning is effective.

The optimized model experimented with different masking strategies. The results showed that the model performed best with a masking strategy of patch size of 64 and masked ratio of 0.5 after transfer learning optimization. Therefore, this article uses this masking strategy to train the augmented data set. The final model obtained is SL-DDBD. In this section, we will compare the accuracy of the baseline, the optimized model trained with original data (Improved), the model trained with the ViT as the encoder, and the Improved+DA model at different epochs, as shown in Fig. 10. The Improved+DA model has the fastest convergence speed and the second is the ViT model. However, the accuracy of the ViT model at the end is

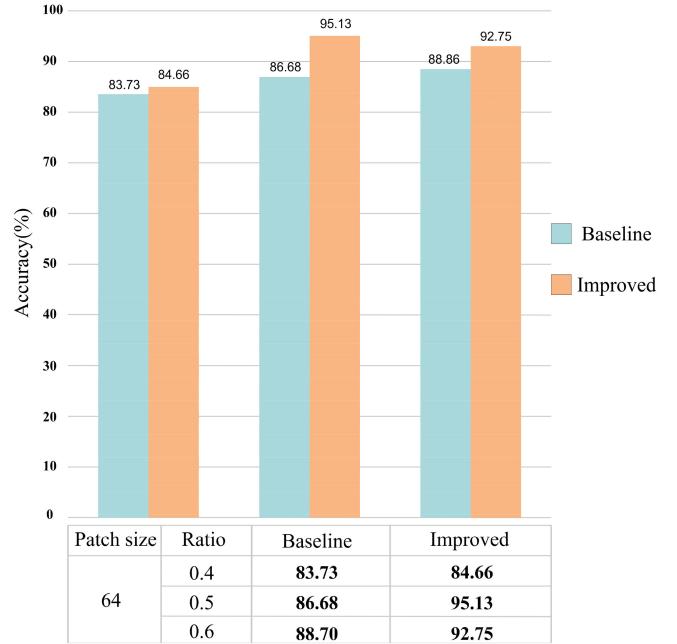


Fig. 9. Comparison of the accuracy of the baseline and improved models with a fixed patch size of 64 and different ratio values.

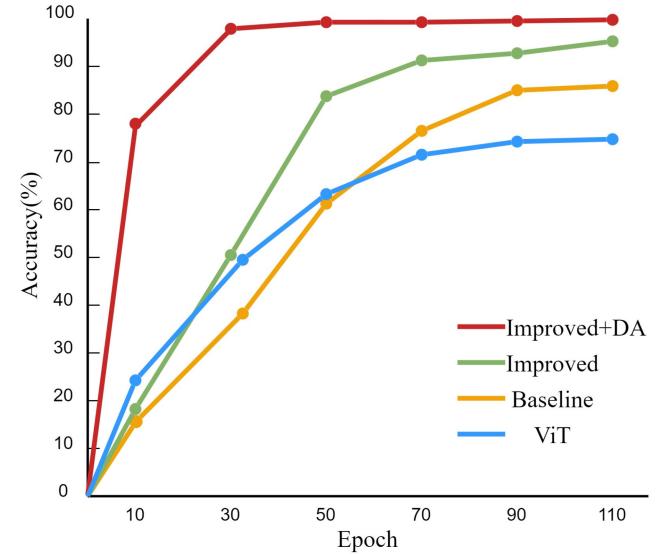


Fig. 10. Accuracy rate comparison of different models.

the lowest, only 74.35%. It is obvious that the Improved+DA model has reached a high accuracy of 78% at epoch 10. At the same epoch, the accuracy of other models is basically kept between 15%–25%. The accuracy of the Improved model is higher than the baseline at all epochs, which again proves the improvement of the optimization work on the model accuracy. The Improved+DA model not only converges quickly but also outperforms other models at each epoch. The accuracy of the model is 99.60%.

At the same time, in this experiment, the baseline model and our proposed model SL-DDBD are compared. The performance of the two models is compared in eight metrics: 1) Acc; 2) Ws; 3) GFLOPS; 4) Params; 5) FPS; 6) precision; 7) recall; and 8) F1-score. As shown in Table V, Fig. 11,

TABLE V
COMPREHENSIVE PERFORMANCE OF SL-DDBD AND BASELINE

Model	Acc (%)	W _S /MB	GFLOPS	Parameters	V/FPS
Baseline	84.92	994.7	15.26	86753474	29
SL-DDBD	99.60	316.3	4.49	27527044	56

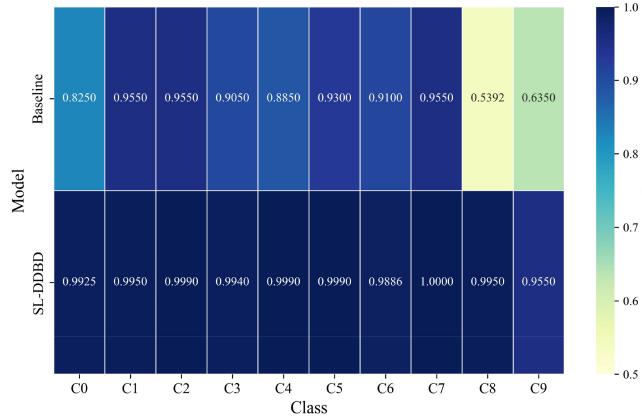


Fig. 11. Heat map comparison of precision for the baseline model and SL-DDBD.

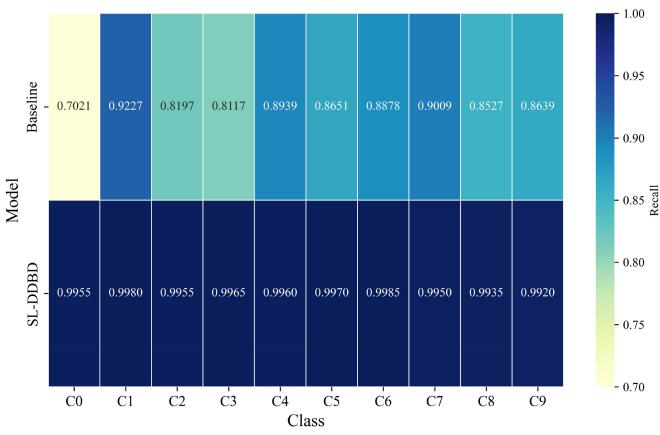


Fig. 12. Heat map comparison of recall for the baseline model and SL-DDBD.

12, and 13. The confusion matrix of the SL-DDBD model is shown in Fig. 14. After the optimization work, our model not only significantly improved the accuracy but also reduced the overall parameters and computation.

The accuracy of the proposed model SL-DDBD reaches the highest level of 99.60%, which is an improvement of 14.68% compared to the baseline model's accuracy of 84.92%. SL-DDBD has shown excellent results in metrics of F1-score, precision, and recall. Especially for categories C8 and C9, the original baseline recognition accuracy was low. However, SL-DDBD significantly improved this situation, maintaining high values for precision and F1-score across all categories. The FPS of the model has also improved after lightweight optimization, from the original 29 FPS to 56 FPS, which is nearly a twofold increase. This not only ensures fast detection but also meets the real-time monitoring requirements. The

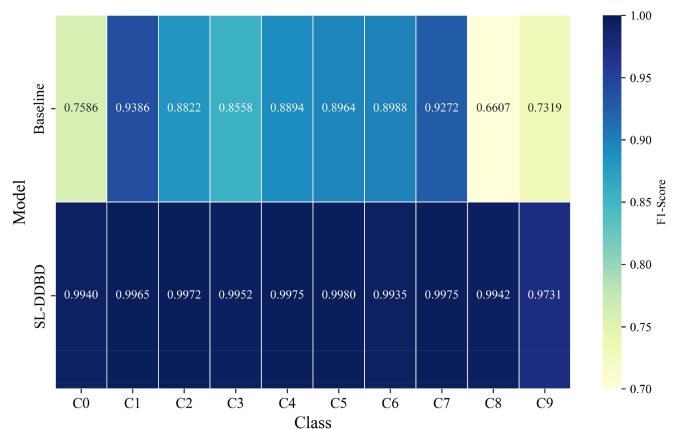


Fig. 13. Heat map comparison of F1-scores for the baseline model and SL-DDBD.

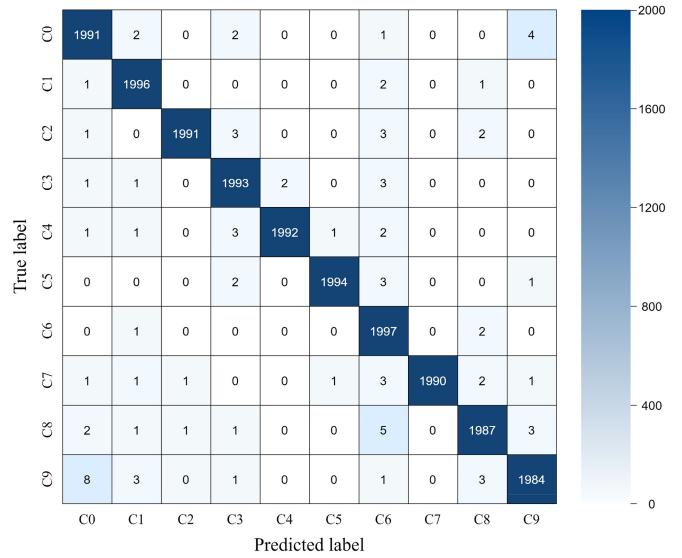


Fig. 14. Confusion matrix of SL-DDBD.

detection performance of the model has reached an acceptable level in terms of both accuracy and detection speed.

Since the self-supervised learning framework was originally designed to solve large-scale computer vision tasks and used a common encoder structure, the model itself is not very lightweight in terms of parameters and weight file size. However, through lightweight optimization in the transfer learning section, the parameters of the model were reduced from 86.7 million to 27.5 million, a decrease of 68.2%. The size of the weight file after training was reduced from 994.7 MB to 316.3 MB. GFLOPS decreased from 15.26 to 4.49, which greatly reduces the difficulty in engineering deployment and hardware requirements.

TABLE VI
COMPARISON RESULTS WITH SUPERVISED MODELS

Model	Parameters(in Millions)	Acc(%)
Alexnet+Softmax[53]	63.2	96.8
Alexnet+Tripletloss[53]	63.2	98.6
D-HCNN[26]	0.76	99.87
HCF[54]	>72.3	96.74
OLCMNet[21]	-	89.53
VGG-GAP[55]	140	98.7
Ensenble VGG-16 and VGG-16[56]	>140	92.6
Driver-net[56]	-	95.0
Vanilla CNN with data augmentation[57]	26.05	97.05
InceptionV3+Xception-50+Xception+VGG-19[58]	214.3	97.00
SL-DDBD	27.5	99.60

“-” No data available

In the confusion matrix, the overall classification of the model is excellent, but it also shows a small number of misclassifications. For example, C0 “Normal driving” and C9 “Talking to passenger” are confused with each other. The main reason for the classification error is the high similarity of the images between the two categories. Drivers occasionally turn their heads during safe driving to observe the front view or to look at the rearview mirror. However, this may be misjudged as talking with passengers. Similarly, in some images of C9 “Talking to passenger,” drivers do not obviously turn their heads to talk to passengers on the right, which is also easily misjudged as “Normal driving.”

E. Comparison With Previous Results

To validate the detection performance of the proposed method, we compared our best model, the “SL-DDBD” model, with previous state-of-the-art supervised and unsupervised learning methods.

The comparison results between the proposed method and supervised learning models are shown in Table VI. The main evaluation metrics are accuracy and parameters, which are used to compare and analyze the model’s accuracy and size.

Literature [53] and [55] both use earlier CNN networks and improve the accuracy through some improvements. Their accuracy has not reached a high value. At the same time, due to the early network, the feature learning ability is poor, which leads to larger network parameters. Models like Ensemble VGG-16 and VGG-16 [55] have already exceeded 140 million in terms of parameters, but the accuracy is still a low value of 92.6%. The method in literature [58] is a fusion of multiple detection models, with parameters reaching 214.3 million, and the accuracy is still 97%. Similarly, the HCF [54] method, which is based on the fusion of multiclass CNN models, reduces the parameters to the lowest value of 72.3 million, but the accuracy is still not high. Recently, the D-HCNN [26] method has achieved high accuracy, exceeding 99%, by decreasing the convolutional kernel size to lightweight. Due to the large data set of self-supervised training and the encoder of the Swin Transformer, the parameters of the basic model are larger compared to CNN models. However, the proposed method in our

work is based on self-supervised learning, which is a particular advantage. The cost of supervised learning training will be significantly reduced, and it has strong generalization capabilities. Transfer learning only requires a small data set and performs well in more detailed downstream tasks. Moreover, we have conducted lightweight work and verified its effectiveness through experiments. Finally, the model’s parameters can be reduced to 27.5 million while achieving a recognition accuracy of 99.6%. Compared to most of the supervised learning-based CNN models in the table, the recognition accuracy is higher and the parameters are smaller.

Currently, there are only a few unsupervised learning methods available in the field of driver distraction behavior recognition. This indicates the need for further exploration and research in driver behavior detection. This article compares the proposed method with recent unsupervised learning methods for driver distraction behavior detection [48], and presents the comparison results in Table VII.

The SimCLR is a classic unsupervised learning method, with a ResNet50 backbone achieving 94.32% accuracy. The Simsiam method, also with a ResNet50 backbone, has an accuracy of 86.29%. Through the process of model refinement, our approach has surpassed the accuracy of both Simsiam and SimCLR, achieving a high-recognition rate of 95.13%. The baseline model’s recognition performance is only 1.37% lower than that of Simsiam. However, it is worth noting that, unlike the models presented in this work which were trained for 110 epochs, all other models underwent a significantly larger number of training epochs, reaching up to 400. In the SimCLR unsupervised learning method, a larger batch size provides more negative examples to facilitate convergence. Longer training epochs also provide more negative examples and can significantly improve results. All models in this work have only 110 training epochs. MIM is a new “prediction” type of the self-supervised learning method. This method has good representation learning ability supported by pretraining, reducing the cost of transfer learning for driver distraction behavior detection. The UDL model proposed in [48] is an improvement on the Simsiam structure and is also a “constructive” category unsupervised learning method like SimCLR. With the main network using RepMLP-Res50, the final accuracy of the UDL model is 98.61%. The accuracy of the SL-DDBD model in this article is 99.60%, which is higher than the recognition accuracy of the all above models.

V. CONCLUSION

This article explores the introduction of the MIM self-supervised learning method in the task of driver distraction behavior detection. Image masking strategy is used for pre-training on a large number of unlabeled data sets. In order to better integrate unsupervised learning and downstream tasks, transfer learning is carried out on the driver distraction behavior data set. Lightweighting and accuracy optimization work has been done in transfer learning.

- 1) The original encoder was improved through a reconfiguration of the number of Swin Transformer blocks in

TABLE VII
COMPARISON RESULTS WITH UNSUPERVISED MODELS

Method	Backbone	Batch-size	Epoch	Acc(%)
Simsiam	ResNet50	32	400	86.29
SimCLR	ResNet50	32	400	94.32
UDL	RepMLP-Res50	32	400	98.61
Baseline	Swin Transformer	32	110	84.92
Improved	Swin Transformer	32	110	95.13
SL-DDBD	Swin Transformer	32	110	99.60

stage 3. The task detection accuracy was improved while reducing the complexity of the encoder network.

- 2) For each stage of the encoder, a new distribution of the number of W-MSA and SW-MSA detection heads was made. The number of feature transfers was reduced and attention was increased on key feature information.
- 3) In transfer learning, the impact of MIM strategy on downstream tasks was considered and a comprehensive comparison experiment was designed. The best masking strategy was selected.
- 4) We used a multiclass data augmentation strategy to simulate real-world scenarios to expand data set. This further improved the model's generalization ability in complex scenarios. SL-DDBD achieves 99.60% accuracy on the large-scale driver distraction behavior data set State-Farm.

In future work, interesting work is to try using multisource information fusion methods for driver distraction behavior detection. we will predict the driver's eye focus position. The predicted eye focus position and the action behavior recognition results are combined to achieve more accurate driver distraction recognition. In addition, we will consider using model pruning methods to further do lightweight research on the self-supervised driver distraction behavior detection model. The model quantifies the weights during transfer learning training, and then deletes low-weight parts according to certain standards. Deploying the model on mobile devices for fast and real-time detection of driver distraction.

REFERENCES

- [1] "Global status report on road safety." Jan. 2018. [Online]. Available: <https://www.who.int/publications/item/global-status-report-on-road-safety-2018>
- [2] "Overview of the 2019 crash investigation sampling system," *Traffic Saf. Facts Res. Note*, vol. 2020, pp. 1–14, Dec. 2020.
- [3] L. Chen et al., "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1046–1056, Feb. 2023.
- [4] D. Wakabayashi. "Self-driving uber car kills pedestrian in arizona where robots room." Sep. 2019. [Online]. Available: <https://www.nytimes.com/2018/03/19/technology/uber-driverlessfatality.html>
- [5] 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.*, vol. 8, no. 1, pp. 37–47, Jan. 2023.
- [6] C. Ahlström, G. Georgoulas, and K. Kircher, "Towards a context-dependent multi-buffer driver distraction detection algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4778–4790, May 2022.
- [7] C. Huang, H. Huang, J. Zhang, P. Hang, Z. Hu, and C. Lv, "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.
- [8] Y. Xing, C. Lv, Y. Liu, Y. Zhao, D. Cao, and S. Kawahara, "Hybrid-learning-based driver steering intention prediction using neuromuscular dynamics," *IEEE Trans. Ind. Electron.*, vol. 69, no. 2, pp. 1750–1761, Feb. 2022.
- [9] "Distracted driving." Accessed: Jan. 12, 2022. [Online]. Available: https://www.cdc.gov/motorvehiclesafety/distracted_driving/
- [10] N. Zangi, R. Srour-Zreik, D. Ridel, H. Chasidim, and A. Borowsky, "Driver distraction and its effects on partially automated driving performance: A driving simulator study among young-experienced drivers," *Accid. Anal. Prevent.*, vol. 166, Mar. 2022, Art. no. 106565.
- [11] P. Hang, C. Lv, Y. Xing, C. Huang and Z. Hu, "Human-like decision making for autonomous driving: A noncooperative game theoretic approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2076–2087, Apr. 2021.
- [12] S. Jha and C. Busso, "Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 59–72, Jan. 2023.
- [13] K. Roy, "Unsupervised sparse, nonnegative, low rank dictionary learning for detection of driver cell phone usage," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18200–18209, Oct. 2022.
- [14] 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.*, vol. 23, no. 10, pp. 19309–19322, Oct. 2022.
- [15] D. He, B. Dommez, C. C. Liu, and K. N. Plataniotis, "High cognitive load assessment in drivers through wireless electroencephalography and the validation of a modified N-back task," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 4, pp. 362–371, Aug. 2019.
- [16] C. Fan, Y. Peng, S. Peng, H. Zhang, Y. Wu, and S. Kwong, "Detection of train driver fatigue and distraction based on forehead EEG: A time-series ensemble learning method," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13559–13569, Aug. 2022.
- [17] X. Wang, R. Xu, S. Zhang, Y. Zhuang, and Y. Wang, "Driver distraction detection based on vehicle dynamics using naturalistic driving data," *Transp. Res. C Emerg. Technol.*, vol. 136, Mar. 2022, Art. no. 103561.
- [18] P. Papantoniou, E. Papadimitriou, and G. Yannis, "Review of driving performance parameters critical for distracted driving research," *Transp. Res. Proc.*, vol. 25, pp. 1796–1805, Jan. 2017.
- [19] J. Hansen, C. Busso, Y. Zheng, and A. Sathyaranayana, "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.
- [20] L. Jiang, W. Xie, D. Zhang, and T. Gu, "Smart diagnosis: Deep learning boosted driver inattention detection and abnormal driving prediction," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4076–4089, Mar. 2022.
- [21] 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.
- [22] B. Wagner, F. Taffner, S. Karaca, and L. Karge, "Vision based detection of driver cell phone usage and food consumption," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4257–4266, May 2022.
- [23] T. Huang, R. Fu, Y. Chen, and Q. Sun, "Real-time driver behavior detection based on deep deformable inverted residual network with an attention mechanism for human-vehicle co-driving system," *IEEE Trans. Intell. Veh.*, vol. 71, no. 12, pp. 12475–12488, Dec. 2022.
- [24] R. Chandra, A. Bera, and D. Manocha, "Using graph-theoretic machine learning to predict human driver behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2572–2585, Mar. 2022.
- [25] M. Tan et al., "Bidirectional posture-appearance interaction network for driver behavior recognition," *IEEE Trans. Intell. Trans. Syst.*, vol. 23, no. 8, pp. 13242–13254, Aug. 2022.

- [26] 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.
- [27] D. Li, Y. Wang, and W. Xu, "A deep multichannel network model for driving behavior risk classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 1204–1219, Jan. 2023.
- [28] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 9992–10002.
- [29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Informat. Process. Syst.*, 2017, pp. 5998–6008.
- [30] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 6000–6010.
- [31] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.
- [32] Z. Hu, Y. Zhang, Q. Li, and C. Lv, "A Novel heterogeneous network for modeling driver attention with multi-level visual content," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24343–24354, Dec. 2022.
- [33] Y. Xing, C. Lv, D. Cao, and E. Velenis, "Multi-scale driver behavior modeling based on deep spatial-temporal representation for intelligent vehicles," *Transp. Res. C Emerg. Technol.*, vol. 130, Sep. 2021, Art. no. 103288.
- [34] A. Behera, Z. Wharton, A. Keidel, and B. Debnath, "Deep CNN body pose and body-object interaction features for drivers' activity monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2874–2881, Mar. 2022.
- [35] T. Huang and R. Fu, "Driver distraction detection based on the true driver's focus of attention," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19374–19386, Oct. 2022.
- [36] T. Li, Y. Zhang, Q. Li, and T. Zhang, "AB-DLM: An improved deep learning model based on attention mechanism and BiFPN for driver distraction behavior detection," *IEEE Access*, vol. 10, pp. 83138–83151, 2022.
- [37] 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.
- [38] D. Liu, T. Yamasaki, Y. Wang, K. Mase, and J. Kato, "Toward extremely lightweight distracted driver recognition with distillation-based neural architecture search and knowledge transfer," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 764–777, Jan. 2023.
- [39] W. Li, J. Wang, T. Ren, F. Li, J. Zhang, and Z. Wu, "Learning accurate, speedy, lightweight CNNs via instance-specific multi-teacher knowledge distillation for distracted driver posture identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17922–17935, Oct. 2022.
- [40] E. Shang, H. Liu, Z. Yang, J. Du, and Y. Ge, "FedBiKD: Federated bidirectional knowledge distillation for distracted driving detection," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11643–11654, Jul. 2023.
- [41] H. Wang et al., "FPT: Fine-grained detection of driver distraction based on the feature pyramid vision transformer," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 1594–1608, Feb. 2023.
- [42] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.
- [43] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.
- [44] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.
- [45] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.
- [46] X. Chen et al., "Context autoencoder for self-supervised representation learning," 2022, *arXiv:2202.03026*.
- [47] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14668–14678.
- [48] B. Li et al., "A new unsupervised deep learning algorithm for fine-grained detection of driver distraction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19272–19284, Oct. 2022.
- [49] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
- [50] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2022, *arXiv:1710.09412*.
- [51] "State farm distracted driver detection." Jun. 2017. [Online]. Available: <https://www.kaggle.com/c/state-farm-distracted-driverdetection>.
- [52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [53] O. D. Okon and L. Meng, "Detecting distracted driving with deep learning," in *Proc. Int. Conf. Interact. Collaborat. Robot.*, 2017, pp. 170–179.
- [54] C. Huang, X. Wang, J. Cao, S. Wang, and Y. Zhang, "HCF: A hybrid CNN framework for behavior detection of distracted drivers," *IEEE Access*, vol. 8, pp. 109335–109349, 2020.
- [55] B. Zhang, *Apply and Compare Different Classical Image Classification Method: Detect Distracted Driver*, Stanford Univ., Stanford, CA, USA, 2016.
- [56] M. S. Majdi, S. Ram, J. T. Gill, and J. J. Rodriguez, "Drive-Net: Convolutional network for driver distraction detection," in *Proc. IEEE Southwest Symp. Image Anal. Interpret. (SSIAI)*, Apr. 2018, pp. 1–4.
- [57] A. V. Jamsheed, B. Janet, and U. S. Reddy, "Real time detection of driver distraction using CNN," in *Proc. 3rd Int. Conf. Smart Syst. Invent. Technol. (ICSSIT)*, Aug. 2020, pp. 185–191.
- [58] K. R. Dhakate and R. Dash, "Distracted driver detection using stacking ensemble," in *Proc. Int. Students Conf. Elect. Electron. Comput. Sci. (SCEECS)*, Feb. 2020, pp. 1–5.

Yingzhi Zhang (Graduate Student Member, IEEE) was born in Laiwu, Shandong, China, in 1998. He received the B.Eng. degree in rail traffic signal and control from Jiangsu Normal University, Xuzhou, China, in 2021. He is currently pursuing the master's degree in electrical engineering with Lanzhou Jiaotong University, Lanzhou, China.

His research interests include driver behavior, computer vision, and machine learning.



Taiguo Li received the B.Eng. degree in computer science and technology and the M.Eng. degree in computer software and theory from Wuhan University, Wuhan, China, in 2008 and 2010, respectively, and the Ph.D. degree in electronic science and technology from Lanzhou Institute of Physics, China Academy of Space Technology, Beijing, China, in 2017.

He is currently working as an Assistant Professor with Lanzhou Jiaotong University, Lanzhou, China. His research interests include electronic science and technology, computer vision, and driver behavior detection.



Chao Li received the bachelor's degree in rehabilitation therapy from Shaanxi University of Chinese Medicine, Xianyang, China, in 2010.

He is currently a Medical Staff of Shaanxi Kangfu Hospital, Xi'an, China. His research interests include human factor, mental attention, and exercise therapy rehabilitation.



Xinghong Zhou was born in Chenzhou, Hunan, China, in 1999. He received the B.Eng. degree in rail traffic signal and control from Shaoyang University, Shaoyang, China, in 2021. He is currently pursuing the master's degree in transportation with Lanzhou Jiaotong University, Lanzhou, China.

His research interests include fault diagnosis, computer vision, and driver fatigue detection.

