ViT-DD: Multi-Task Vision Transformer for Semi-Supervised Driver Distraction Detection

Yunsheng Ma[*]**, Ziran Wang**
Purdue University College of Engineering
West Lafayette, IN 47907
{ma801, ziran}@purdue.edu

Abstract

Driver distraction detection is an important computer vision problem that can play a crucial role in enhancing traffic safety and reducing traffic accidents. This paper proposes a novel semi-supervised method for detecting driver distractions based on Vision Transformer (ViT). Specifically, a multi-modal Vision Transformer (ViT-DD) is developed that makes use of inductive information contained in training signals of distraction detection as well as driver emotion recognition. Further, a self-learning algorithm is designed to include driver data without emotion labels into the multi-task training of ViT-DD. Extensive experiments conducted on the SFDDD and AUCDD datasets demonstrate that the proposed ViT-DD outperforms the best state-of-the-art approaches for driver distraction detection by 6.5% and 0.9%, respectively.

1 Introduction

According to National Highway Traffic Safety Administration (NHTSA), there were 38,824 people killed in motor vehicle crashes on U.S. roadways during 2020. Among these cases, 3,142 (or 8.1%) are distraction-affected crashes, i.e., a crash involving at least one driver who was distracted. Distracted driving is defined by NHTSA as any activity that diverts attention away from safe driving, such as talking or texting on a cell phone, eating and drinking, chatting with others in the car, and fiddling with the audio, entertainment, or navigation system [1].

During the past decade, rapid development has been witnessed worldwide in the intelligent vehicle technology, where advancements in perception, communication, and computation introduced numerous emerging applications on intelligent vehicles. As a core element of intelligent vehicles, driving automation systems, such as Advanced Driver-Assistance Systems (ADAS) and Automated Driving Systems (ADS), have been designed to support human drivers either by providing warnings to reduce risk exposure, or by assisting the vehicle actuation to relieve drivers' burden on some of the driving tasks. When functioning, these systems can help the driver safely navigate the vehicle through tricky traffic scenarios when him/her is distracted by some other tasks [2].

However, a driver can also over-trust the driving automation system, especially when the system is categorized as SAE Level 3 (i.e., conditional driving automation): When the automated driving features are engaged, the driver is allowed to take his/her hands off the steering wheel and feet off the pedals, but he/she needs to stay alert and get ready to take over the driving task when the system requests [3]. Due to the human nature, the attention from the driver on road conditions can get diminished when he/she is not in charge of driving, and the involvement of distracted behaviors can decrease driver's capability of taking over, which in turn leads to traffic accidents.

[*] Corresponding author

Machine Learning for Autonomous Driving Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, USA.

It can be envisioned in the future transportation systems that intelligent vehicles can detect and identify driver distractions, then warn the driver against them or take precautionary measures. Therefore, in this paper, a multi-modal Vision Transformer (termed ViT-DD) is proposed to exploit inductive information contained in the training signals of both emotion recognition and distraction detection, along with a novel pseudo-labeled multi-task training algorithm which leverages the knowledge in an independent emotion recognition teacher model to train a student ViT-DD.

In summary, the contributions of this paper are threefold:

- To the best of the authors' knowledge, this is the first paper to explore the detection of driver distractions using a pure Transformer-based architecture.
- A multi-modal Vision Transformer, i.e. ViT-DD, is developed for driver distraction detection, where a novel semi-supervised learning approach is proposed to include driver data without emotion labels in the multi-task training of ViT-DD.
- Extensive experiments on the SFDDD [4] and AUCDD [5] datasets are conducted, and the results demonstrate the superiority of the proposed methodologies compared to the state-of-the-art approaches.

2 Background

2.1 Vision Transformer

Since AlexNet [6], convolutional neural networks (CNNs) [7] have been the dominant methodology for learning visual representations of images in computer vision (CV) [8, 9]. Vision Transformer (ViT) [10], on the other hand, has recently achieved state-of-the-art performances on a variety of CV tasks and garnered significant interest from the CV community. For instance, Can et al. [11] propose employing Transformer models for traffic scene understanding tasks.

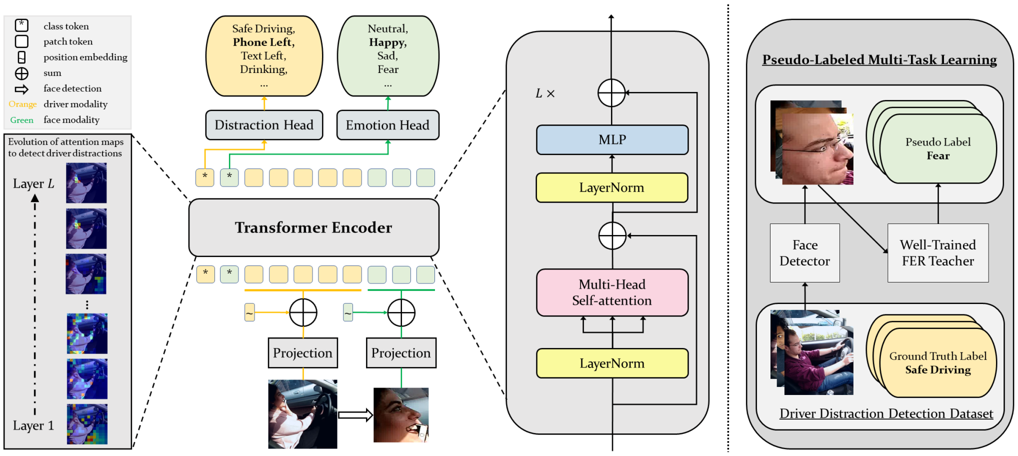
ViT seldom employs convolution kernels (i.e. the core of CNNs). Instead, it relies on the self-attention mechanism [12] to provide context information for input visual tokens, which is inspired by tasks in natural language processing. In addition to the initial patch extraction process, ViT does not introduce image-specific inductive biases into its architecture.

Despite the success of ViT, there has been few works to apply the Transformer architecture to the task of driver distraction detection. In this paper, Vision Transformer is adopted as the backbone network, and it is extended so that it can receive multi-modal input images and be used in multi-task learning setting.

2.2 Multi-Task Learning and Self-Training

Multi-Task Learning (MTL) is an inductive transfer mechanism with the primary objective of enhancing generalization performance[13]. Learning one task at a time is the standard for machine learning. However, Caruana[13] contends that this strategy is sometimes ineffective, since it disregards a potentially rich source of information accessible in many real-world problems, i.e., the information contained in the training signals of other tasks drawn from the same domain. If the tasks can share what they learn, it may be preferable to require the learner to learn many capabilities simultaneously. In computer vision, a popular method for MTL is to employ a single encoder to learn a shared representation, followed by numerous task-specific decoders [14, 15]. In this paper, a similar strategy is employed by training one main backbone model together with several small task-specific heads.

Self-training is an approach for incorporating unlabeled data into a supervised learning task [16, 17, 18]. It is one of the earliest semi-supervised learning approaches, which generates pseudo labels for unlabeled data using a supervised model. Recently, Ghiasi et al.[15] proposes multi-task self-training (MuST), an approach for generating generalized visual representations using multi-task learning with pseudo labels. This method differs from the approach presented in this paper in that the multi-task learning strategy is employed for both output and various input modalities.



class token
patch token
position embedding
sum
face detection
Orange driver modality
Green face modality
Evolution of attention maps to detect driver distractions
Layer L
Layer 1
Safe Driving, Phone Left, Text Left, Drinking, ...
Neutral, Happy, Sad, Fear ...
Distraction Head
Emotion Head
Transformer Encoder
Projection
Projection
L ×
MLP
LayerNorm
Multi-Head Self-attention
LayerNorm
Pseudo-Labeled Multi-Task Learning
Pseudo Label Fear
Face Detector
Well-Trained FER Teacher
Ground Truth Label Safe Driving
Driver Distraction Detection Dataset

Figure 1: **(Left) The framework of the proposed ViT-DD:** First, a face detector is applied to the input signal from an in-cabin camera to acquire the driver's facial area. Then, the driver and face images are divided into patches and independently embedded into visual tokens. Next, the driver and face embeddings are added with their respective position embeddings, and the resulting sequence is concatenated. In addition, tokens representing distractions and emotions are prepended. The sequence of class and visual tokens are then iteratively updated through L Transformer layers. The class tokens from the final sequence are used to recognize the driver's distraction and emotion states through their corresponding MLP heads. **(Right) Pseudo-labeled Multi-Task Learning:** A well-trained Facial Expression Recognition Teacher ViT is employed to label the unlabeled drivers' face images in order to create a multi-task driver dataset. The dataset containing both ground-truth distraction labels and pseudo emotion labels is then applied to train a student ViT-DD model with multi-task learning.

2.3 Facial Expression Recognition and Driver Distraction Detection

Facial expression recognition (FER) is an image classification problem that recognizes the emotion state of individuals [19, 20]. AffectNet-7 [21] is currently the largest publicly available FER dataset, which comprises images labeled with Ekman's six fundamental emotions [22], namely *happy, sad, surprise, fear, disgust,* and *anger*, plus an additional *neutral* category. In this paper, FER is employed to drivers' face images to evaluate his or her emotion state, therefore acquiring additional information to detect driver distractions through multi-task learning.

3 Methodology

In this section, a novel multi-task ViT for semi-supervised driver distraction detection is proposed, where the overall framework is shown in Figure 1. Specifically, ViT-DD has two input modalities, i.e. driver and face, to exploit information contained in the training signals of both distraction detection and emotion recognition. The input images from both modalities are separated into patches, linearly projected to fixed-dimensional visual tokens, and encoded using a Transformer encoder. Task-specific classification heads are applied to the output sequence of the Transformer encoder to generate the prediction results. The training of ViT-DD is conducted through a novel multi-task multi-modal self-training technique.

3.1 Model Overview

The backbone of ViT-DD is a ViT [10], with different patch projection layers applied to each input modality (driver and face). Specially, the input space for each modality is defined by $\mathcal{X}^{(i)}$, where $i \in \{0, 1\}$. The input image $\mathbf{x}^{(i)} \in \mathcal{X}^{(i)} \subseteq \mathbb{R}^{C \times H_i \times W_i}$ from modality i is sliced into patches and then flattened to $\mathbf{v}^{(i)} \in \mathbb{R}^{N_i \times (P^2 \cdot C)}$, where (P, P) is the patch size, C is the number of channels of the input image, and $N_i = H_i W_i / P^2$ is the number of patches for each modality. Next, the flattened patches are linear projected to D dimensional tokens with the projection matrix $E^{(i)} \in \mathbb{R}^{(P^2 \cdot C) \times D}$, followed by position embedding $E_{\text{pos}} \in \mathbb{R}^{N_i \times D}$. Additionally, class tokens ($\mathbf{t}_{\text{class}}^{(i)}$) for each modality with a learnable embedding are prepended to the beginning of the input sequence. All input tokens are then concatenated into a combined sequence $\mathbf{z}_0$ (Eq.2) and sent to the same Transformer Encoder as input.

$$\begin{aligned}
\bar{\mathbf{x}}^{(i)} &= [\mathbf{t}_1^{(i)} E^{(i)}; \cdots ; \mathbf{t}_{N_i}^{(i)} E^{(i)}] + E_{\text{pos}}^{(i)}, && i = 0, 1 && (1)\\
z^0 &= [\mathbf{t}_{\text{class}}^{(0)}; \mathbf{t}_{\text{class}}^{(1)}; \bar{\mathbf{x}}^{(0)}, \bar{\mathbf{x}}^{(1)}] && && (2)\\
\mathbf{z}'_\ell &= \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, && \ell = 1 \dots L && (3)\\
\mathbf{z}_\ell &= \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, && \ell = 1 \dots L && (4)\\
\mathbf{y}^{(i)} &= \text{LN}(\mathbf{z}_L^i), && i = 0, 1 && (5)
\end{aligned}$$

The Transformer encoder then learns the fused driver behavior and emotion representation by stacking L Transformer blocks . A Multilayer Perceptron (MLP) module and a Multihead Self-attention (MSA) module are included in each block. Additionally, LayerNorm (LN) [23] is also adopted prior to each module. Self-attention (SA) is the key component of Transformer blocks, in which, the input vector $\mathbf{z}_{\ell-1}$ is first transformed into three separate vectors: the query vector $\mathbf{q}$, the key vector $\mathbf{k}$, and the value vector $\mathbf{v}$, all of the same dimension $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^D$. After that, the attention scores are constructed by the following function:

$$\text{SA}(\mathbf{z}_{\ell-1}) = \text{softmax}\left(\frac{W_Q\mathbf{q} \cdot (W_K\mathbf{k})^\top}{\sqrt{D_H}}\right)(W_V\mathbf{v}) \tag{6}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D_H \times D}$ are learnable parameters of three linear projections and $D_H = D$.

MSA is an extension of SA with H self-attention heads, which can be formulated as:

$$\text{MSA}(\mathbf{z}_{\ell-1}) = W_P\,[\text{SA}_1(\mathbf{z}_{\ell-1}); \cdots ; \text{SA}_H(\mathbf{z}_{\ell-1})] \tag{7}$$

where $W_P \in \mathbb{R}^{D \times (H \cdot D_H)}$, and D_H is typically set to D/H. To obtain the final multi-task prediction probabilities of the driver's distractions and emotions, the states of the *class* tokens at the output of the Transformer encoder ($\mathbf{z}_L^i$) are fed into the respective classification heads, with the standard cross-entropy loss adopted. The final loss is the weighed sum of the loss of each head.

For all experiments, a ViT-B [10] pretrained on ImageNet[24] with a patch size of 16×16 pixels is employed as the backbone. Specially, the latent vector size D is 768, the patch size P is 16, the layer depth L is 12, and the number of attention heads H is 12.

3.2 Pseudo-Labeled Multi-Task Training

Pseudo labeling offers the benefit of not requiring a large multi-task dataset with all required labels. Having access to a well-trained neural network, that can produce pseudo labels of other domains on the dataset we wish to work on, can be effective. Pseudo labeling is a one-time preprocessing method applicable to RGB datasets of variable size. Compared to the training cost, this phase is computationally inexpensive [14].

The proposed multi-task multi-modal self-training algorithm has four steps. First a teacher ViT is trained on AffectNet-7 [21], a large facial emotion recognition dataset, to enable it to recognize the facial expressions of drivers. Second, RetinaFace[25], a face detector, is used to detect and crop face images in the the driver distraction detection datasets. Next, the FER teacher model is used to label the unlabeled drivers' face images to create a multi-task pseudo-labeled driver dataset. Finally, the driver dataset, which now contains both supervised labels for distraction detection and pseudo labels from the teacher model for emotion recognition, is then employed to train a student ViT-DD model with multi-task multi-modal learning. To manage the situation in which the driver's face cannot be detected, an additional *Non-Face* label is added to the emotion classification task, and in such case, a blank image is fed to the face input.

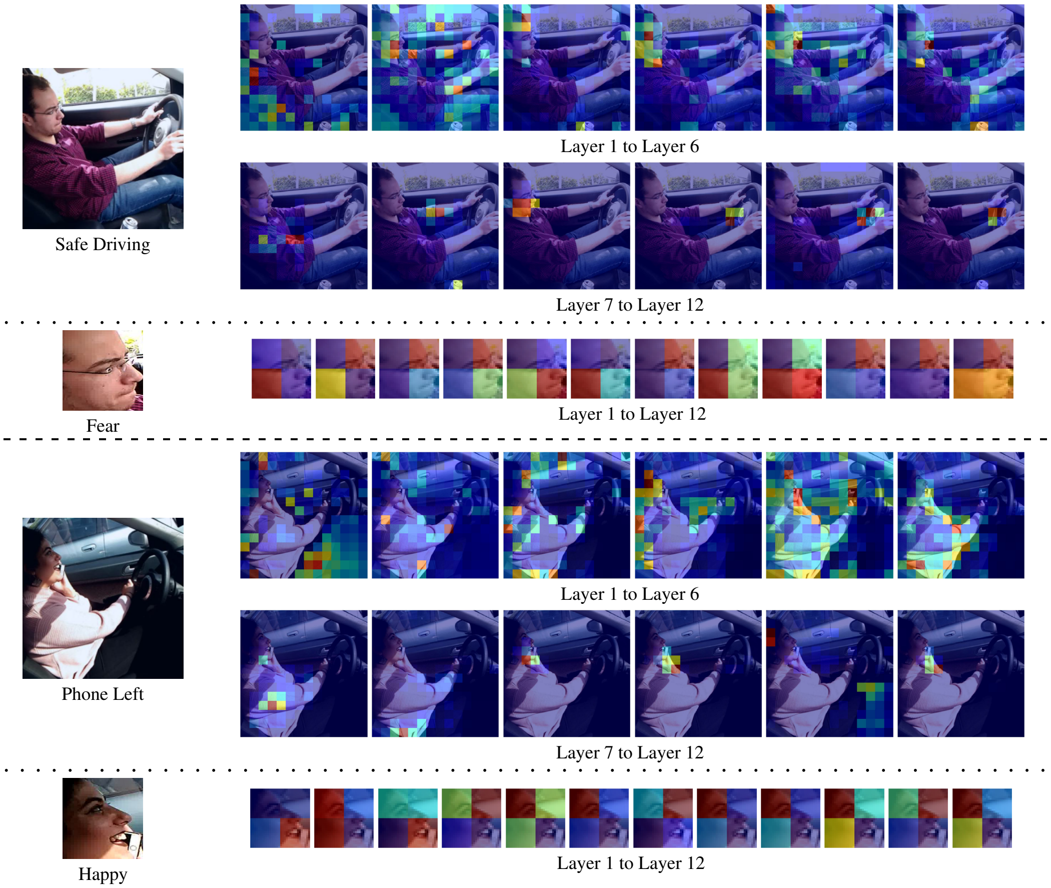


Figure 2: Visualization of the attention maps between the distraction token and all visual tokens in the $L = 12$ Transformer layers of ViT-DD. Colors visualize the attended regions of the model, where red and blue represent regions with high and low attention, respectively.

4 Experiments and Results

In this section, the performance of the proposed ViT-DD model in detecting driver distractions is assessed. The employed benchmarks and baselines are described first. The major results are then reported along with some empirical analysis.

4.1 Benchmarks

The performance of ViT-DD is evaluated on two publicly available distracted driver detection datasets: State Farm Distracted Driver Detection (SFDDD) and the American University in Cairo Distracted Driver Dataset (AUCDD). The SFDDD dataset is comprised of 22,424 labeled images of 26 drivers captured by a constant-placed 2D dashboard camera with 640 × 480 pixels in RGB [4]. The AUCDD dataset contains 10,555 training images and 1123 testing images with resolution of 1920 × 1080 pixels in RGB. It includes data for 44 drivers, 38 of whom are included in the training set and 6 in the test set [5, 26]. Both datasets cover the same ten real-world driving postures: (C0) *Safe Driving*, (C1) *Phone Right*, (C2) *Phone Left*, (C3) *Text Right*, (C4) *Text Left*, (C5) *Adjusting Radio*, (C6) *Drinking*, (C7) *Hair or Makeup*, (C8) *Reaching Behind*, and (C9) *Talking to Passenger*.

There are two commonly used train/test split methods for the SFDDD dataset. One option is to directly split the dataset by images for training or testing, but it will result in a strong correlation between the training and testing data. In particular, it is possible that consecutive video frames are divided into training and testing sets, so it simplifies the problem of distraction detection. The other one is to divide the dataset by drivers such that images of the same driver do not appear in both training and test data.

Table 1: Comparison between the proposed ViT-DD with several state-of-the-art methods. The best method among each setting is highlighted in **bold**. ↓ indicates lower is better. ↑ indicates higher is better. * Results from the original papers. Our method achieve the highest average accuracy for all three experiments, outperforming the state-of-the-art approaches.

Experiment	Method	Accuracy (↑)	NLL (↓)
AUCDD	GA-Weighted Ensemble*[5]	0.9006	0.6400
	ADNet*[27]	0.9022	–
	C-SLSTM*[28]	0.9270	0.2793
	ViT-DD (ours)	**0.9359**	**0.2399**
SFDDD Split-by-Driver	DD-RCNN*[29]	0.8600	**0.3900**
	ViT-DD (ours)	**0.9251**	0.3972
SFDDD Split-by-Image	ViTConv* [30]	0.9790	0.0800
	Inception+ResNet+HRNN* [31]	0.9930	–
	LWANet* [32]	0.9937	0.0260
	ViT-DD (ours)	**0.9963**	**0.0171**

In this paper, results of both split approaches for the SFDDD dataset are reported. For the first split method, $70\%/30\%$ of images are used for training and testing, respectively. While for the second, 18/6 of the total 28 drivers are randomly selected for training/testing. The AUCDD dataset adheres to the original split-by-driver setup.

4.2 Baselines

To compare ViT-DD with the state-of-the-art approaches for distracted driver detection, the following methods are selected as baselines:(1) GA-Weighted Ensemble[5], (2) ADNet[27], (3) C-SLSTM[28], (4) DD-RCNN[29], (5) ViTConv[30], (6) Inception+ResNet+HRNN[31], and (7) LWANet[32]. Note that LWANet is not compared on the AUCDD dataset, as it does not adhere to the split-by-driver setting.

4.3 Implementation Details

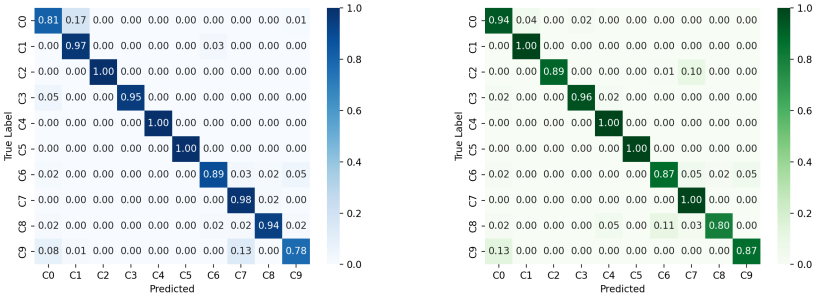
For all experiments, AdamW [33] is adopted as the optimizer with weight decay of 0.1. The base learning rates for SFDDD and AUCDD datasets are set to 0.0003 and 0.0006, respectively. The learning rate is warmed up for 5 epochs, starting with a learning rate of 10^{-6} and decaying to 0 using the cosine scheduler[34]. The input resolution is 224×224 for the driver's images and 32×32 for face images. With a patch size of 16×16, the total number of patches is 200. Simple Random Crop introduced by Touvron et al.[35] with random horizontal flip is employed as the data augmentation strategy. For the SFDDD dataset, 3-Augment introduced in [35] is also applied. Since datasets in this paper are not very large, only the multi-head self-attention layers in the Transformer encoder are fine-tuned, as suggested by Touvron et al.[36]. The ViT-DD model is trained for 20 epochs on 1 NVIDIA A100 GPU with a batch size of 256. Our model is implemented in PyTorch[37].

4.4 Comparison with State-of-the-Art

The performance comparisons between the proposed ViT-DD and the state-of-the-art approaches are shown in Table 1. From the results, we have the following observations:

(1) Splitting-by-image produces a significant correlation between training and testing data. In comparison to other state-of-the-art results on SFDDD with this setting, such as LWANet[32], ViT-DD achieves an accuracy improvement of 0.26%. This demonstrates the excellent fitting capability of ViT-DD.

(2) When adopting a more challenging and realistic split strategy, namely, separate by driver, ViT-DD can respectively obtain 6.5% and 0.9% performance gains over the reported state-of-the-art results, i.e., DD-RCNN[29] on SFDDD and C-SLSTM[28] on AUCDD. This result shows the superior generalization ability of ViT-DD. The performance improvements benefit from the advantages of ViT-DD. First the state-of-the-art ViT is employed as the backbone network, which can provide



True Label	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
C0	0.81	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
C1	0.00	0.97	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00
C2	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C3	0.05	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00
C4	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
C5	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
C6	0.02	0.00	0.00	0.00	0.00	0.00	0.89	0.03	0.02	0.05
C7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.02	0.00
C8	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.94	0.02
C9	0.08	0.01	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.78

(a) Standard ViT

True Label	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
C0	0.94	0.04	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
C1	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C2	0.00	0.00	0.89	0.00	0.00	0.00	0.01	0.10	0.00	0.00
C3	0.02	0.00	0.00	0.96	0.02	0.00	0.00	0.00	0.00	0.00
C4	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
C5	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
C6	0.02	0.00	0.00	0.00	0.00	0.00	0.87	0.05	0.02	0.05
C7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
C8	0.02	0.00	0.00	0.00	0.05	0.00	0.11	0.03	0.80	0.00
C9	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.87

(b) ViT-DD

Figure 3: Confusion matrices of the standard ViT and ViT-DD on AUCDD

high generalization performance if pretrained on a large-scale dataset[38]. Second, the novel multi-task multi-modal self-training method enables ViT-DD to leverage additional inductive information provided by the training signals for recognizing the emotion state of drivers, thereby improving generalization performance.

4.5 Ablation Study

The proposed ViT-DD is trained utilizing a novel multi-task multi-modal self-training procedure. To further validate the effectiveness of this strategy, an ablation study is conducted in comparison to the standard ViT trained with supervised driver distraction detection labels. The average accuracy and NLL on SFDDD split-by-driver and AUCDD datasets are shown in Table 2. The confusion matrices on the AUCDD dataset is shown in Figure 3.

Table 2: Performance comparison between the standard ViT and the proposed ViT-DD.

Dataset	Method	Accuracy (↑)	NLL (↓)
SFDDD	Standard ViT[10]	0.9036	0.5355
	ViT-DD	**0.9251**	**0.3972**
AUCDD	Standard ViT[10]	0.9092	0.2895
	ViT-DD	**0.9359**	**0.2399**

From Table 2, it is clear that for both datasets, ViT-DD performs better. The average accuracy improvements of ViT-DD over the standard ViT are 2.2% and 2.7% on the SFDDD and AUCDD datasets, respectively. This demonstrates that ViT-DD successfully leverages additional sources of information from the emotion recognition to improve the performance of learning on the task of distraction detection. From the confusion matrices, we have the following observations:

(1) ViT-DD performs significantly better in detecting *safe driving* (C0) and *talking to passenger* (C9) with 13% and 9% increases in accuracy, respectively. This is because certain emotion states correlate strongly with these two driving behaviors. Specifically, in most cases, drivers have a neutral emotion when driving safely and tend to be happy when talking to passengers. Standard ViT tends to misclassify *safe driving* as *phone right* (C1). This can be resolved with the support of drivers' emotion information, as talking on the phone corresponds to all kinds of emotion status, not just neutral. Also, in standard ViT, 13% of talking to passenger scenarios are misclassified as *hair or makeup* (C7), compared to 0% in ViT-DD due to the inclusion of emotion information.

(2) However, ViT-DD suffers a performance loss when detecting *phone left* (C2) and *reaching behind* (C8). Specifically, *phone left* is occasionally interpreted as *hair or makeup* (C7), and *reaching behind* is occasionally interpreted as *drinking* (C6). In both of these cases, emotion information may mislead the detection of driver distractions: The driver's emotion state can vary when phoning, so emotion cannot provide useful information for detecting this behavior; For reaching behind, it is difficult to identify the driver's emotion, so the emotion data may not be accurate.

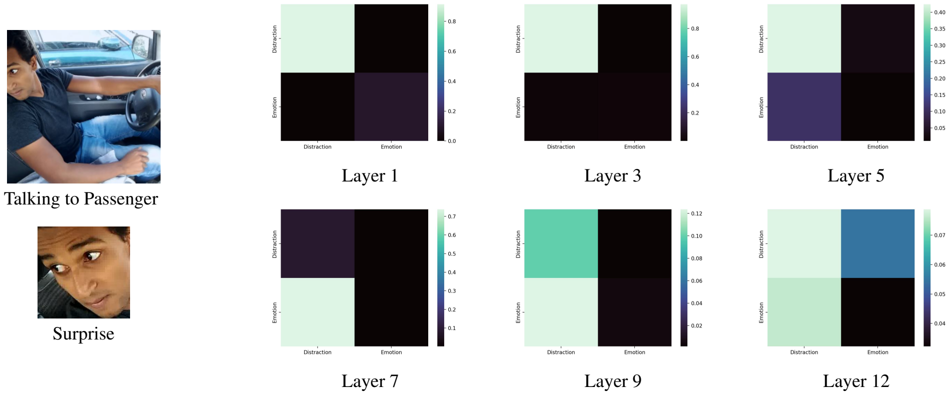


Figure 4: The attention interactions between the distraction token and the emotion token in the 1st, 3rd, 5th, 7th, 9th and 12th Transformer encoder layers of ViT-DD.

(3) It is worth noting that ViT-DD has a much higher detection accuracy on *phone right* (C1) than *phone left* (C2), which is due to the bias present in the dataset. The dash-board camera is positioned in front of the passenger's seat and photographs the driver from the right-hand side. As a result, detection of the phone on the right is much simpler than on the left, since the phone on the right is completely visible.

4.6 Visualization

In order to show the interpretability of our model, the attention maps during inference on the AUCDD dataset are visualized in Figure 2 and Figure 4. Figure 2 shows the interactions between the distraction token and visual tokens of various Transformer encoder layers. The attention scores are used to generate the attention maps. For visualization purposes, the 1D sequence of attention scores is reshaped according to their original spatial positions in the driver or face images.

As seen in Figure 2, as the network becomes deeper, the distraction token gathers more precise local cues rather of the whole driver or face image signals. In the first few layers, the whole in-cabin scene provides interference cues, but a well-trained ViT-DD can gradually concentrate on critical areas of input images. For instance, in the first *safe driving* scenario, the model successfully focuses on the driving wheel region of the driver's image, which is the most informative area of the whole picture. For the second *phone left* scenario, the model effectively pays most attention to the phone region. For both face images, the model attends to the eye region, which is the most distinguishable part of the face for recognizing facial expressions of drivers.

Figure 4 illustrates the attention maps between the distraction token and the emotion token in several self-attention layers. During the initial stages, class tokens show little interaction with one another. As the layer depth increases, both class tokens tend to rely on each other to acquire clues for the final prediction.

5 Conclusion and Future Work

In this paper, a pure Transformer architecture-based method for detecting driver distractions is proposed. The developed ViT-DD trained with the novel pseudo-labeled multi-task learning algorithm can leverage information from emotion recognition to improve the performance of learning on distraction detection. Extensive experiments conducted on SFDDD and AUCDD benchmarks with the challenging split-by-driver setting demonstrate that ViT-DD achieves 6.5% and 0.9% performance improvements as compared to the best state-of-the-art driver distraction detection approaches. As the next step, additional training signals available from in-cabin camera, such as gaze tracking and head pose tracking, can be incorporated into distraction detection or other driver behavior prediction tasks.

References

[1] National Highway Traffic Safety Administration. *Overview of Motor Vehicle Crashes in 2020*. Accessed 12-September-2022. 2022. URL: `https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813266`.

[2] Amir Mukhtar, Likun Xia, and Tong Boon Tang. "Vehicle detection techniques for collision avoidance systems: A review". In: *IEEE transactions on intelligent transportation systems* 16.5 (2015), pp. 2318–2338.

[3] SAE. *SAE Levels of Driving Automation™ Refined for Clarity and International Audience*. Accessed 12-September-2022. 2022. URL: `https://www.sae.org/blog/sae-j3016-update`.

[4] *State Farm Distracted Driver Detection*. Accessed 3-September-2022. 2016. URL: `https://www.kaggle.com/c/state-farm-distracted-driver-detection/overview`.

[5] Hesham M Eraqi et al. "Driver distraction identification with an ensemble of convolutional neural networks". In: *Journal of Advanced Transportation* 2019 (2019).

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[7] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[8] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2015).

[9] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[10] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2020.

[11] Yigit Baran Can et al. "Structured Bird's-Eye-View Traffic Scene Understanding From Onboard Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 15661–15670.

[12] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[13] Rich Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pp. 41–75.

[14] Roman Bachmann et al. "MultiMAE: Multi-modal Multi-task Masked Autoencoders". In: *arXiv preprint arXiv:2204.01678* (2022).

[15] Golnaz Ghiasi et al. "Multi-task self-training for learning general representations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 8856–8865.

[16] Dong-Hyun Lee et al. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2. 2013, p. 896.

[17] Barret Zoph et al. "Rethinking pre-training and self-training". In: *Advances in neural information processing systems* 33 (2020), pp. 3833–3845.

[18] Hieu Pham et al. "Meta pseudo labels". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 11557–11568.

[19] Sicheng Zhao et al. "An end-to-end visual-audio attention network for emotion recognition in user-generated videos". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 303–311.

[20] Aayushi Chaudhari et al. "ViTFER: Facial Emotion Recognition with Vision Transformers". In: *Applied System Innovation* 5.4 (2022), p. 80.

[21] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild". In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.

[22] Paul Ekman. "An argument for basic emotions". In: *Cognition & Emotion* 6.3-4 (1992), pp. 169–200.

[23] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[24] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[25] Jiankang Deng et al. “Retinaface: Single-shot multi-level face localisation in the wild”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5203–5212.

[26] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. “Real-time distracted driver posture classification”. In: *arXiv preprint arXiv:1706.09498* (2017).

[27] Weichu Xiao et al. “Attention-based deep neural network for driver behavior recognition”. In: *Future Generation Computer Systems* 132 (2022), pp. 152–161.

[28] Jimiama Mafeni Mase et al. “A hybrid deep learning approach for driver distraction detection”. In: *2020 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE. 2020, pp. 1–6.

[29] Mingqi Lu, Yaocong Hu, and Xiaobo Lu. “Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals”. In: *Applied Intelligence* 50.4 (2020), pp. 1100–1111.

[30] Yuan Li et al. “Distracted Driving Detection by Combining ViT and CNN”. In: *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE. 2022, pp. 908–913.

[31] Munif Alotaibi and Bandar Alotaibi. “Distracted driver classification using deep learning”. In: *Signal, Image and Video Processing* 14.3 (2020), pp. 617–624.

[32] Yingcheng Lin et al. “A Lightweight Attention-Based Network towards Distracted Driving Behavior Recognition”. In: *Applied Sciences* 12.9 (2022), p. 4191.

[33] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations*. 2018.

[34] Ilya Loshchilov and Frank Hutter. “Sgdr: Stochastic gradient descent with warm restarts”. In: *arXiv preprint arXiv:1608.03983* (2016).

[35] Hugo Touvron, Matthieu Cord, and Hervé Jégou. “Deit iii: Revenge of the vit”. In: *arXiv preprint arXiv:2204.07118* (2022).

[36] Hugo Touvron et al. “Three things everyone should know about Vision Transformers”. In: *arXiv preprint arXiv:2203.09795* (2022).

[37] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[38] Dustin Tran et al. “Plex: Towards reliability using pretrained large model extensions”. In: *arXiv preprint arXiv:2207.07411* (2022).