sustainability

MDPI

Article

Towards Sustainable Safe Driving: A Multimodal Fusion Method for Risk Level Recognition in Distracted Driving Status

Huiqin Chen [1,*], Hao Liu [1], Hailong Chen [1] and Jing Huang [2]

[1] College of Mechanical Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[2] College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, China
* Correspondence: chenhuiqinfj@126.com

Abstract: Precise driving status recognition is a prerequisite for human–vehicle collaborative driving systems towards sustainable road safety. In this study, a simulated driving platform was built to capture multimodal information simultaneously, including vision-modal data representing driver behaviour and sensor-modal data representing vehicle motion. Multisource data are used to quantify the risk of distracted driving status from four levels, safe driving, slight risk, moderate risk, and severe risk, rather than detecting action categories. A multimodal fusion method called vision-sensor fusion transformer (V-SFT) was proposed to incorporate the vision-modal of driver behaviour and sensor-modal data of vehicle motion. Feature concatenation was employed to aggregate representations of different modalities. Then, successive internal interactions were performed to consider the spatiotemporal dependency. Finally, the representations were clipped and mapped into four risk level label spaces. The proposed approach was evaluated under different modality inputs on the collected datasets and compared with some baseline methods. The results showed that V-SFT achieved the best performance with an recognition accuracy of 92.0%. It also indicates that fusing multimodal information effectively improves driving status understanding, and V-SFT extensibility is conducive to integrating more modal data.

Keywords: distracted driving status; vision-sensor fusion transformer; multimodal information; risk level recognition

Citation: Chen, H.; Liu, H.; Chen, H.; Huang, J. Towards Sustainable Safe Driving: A Multimodal Fusion Method for Risk Level Recognition in Distracted Driving Status. *Sustainability* **2023**, *15*, 9661. https://doi.org/10.3390/su15129661

Academic Editors: Kun Gao, Bo Yu, Yang Liu and Jieyu Fan

Received: 23 March 2023
Revised: 12 May 2023
Accepted: 26 May 2023
Published: 16 June 2023

1. Introduction

In the human-machine codriving system, the driver and the intelligent algorithm cooperate and restrict each other to jointly control the vehicle. Accurately identifying the driver's current driving status or risk is an important basis for the intelligent vehicle system to allocate or switch control rights [1]. Many factors can affect a driver's driving status, such as alcohol, fatigue, drugs, negative emotions, and lack of concentration, which may significantly reduce the driver's driving ability and increase potential driving risks. Risky driving not only brings traffic safety hazards [2], but its fuel efficiency may also have a negative impact on environmental protection and sustainability.

Distracted driving and road safety. Distraction is an important cause of risky driving status. The latest data from the National Highway Transportation and Safety Administration (NHTSA) show that approximately 3142 people were killed by distracted driving in 2020, 3119 in 2019, 2628 deaths in 2018, and 3003 deaths in 2017 [3]. Rahman et al. [4] and Sayed et al. [5] focused on studying driving behaviours to precisely anticipate, prevent, and manage road safety programs. Questionnaire measuring [6] and a self-reported survey [7] revealed that risky driving behaviors were associated with traffic safety both directly and indirectly. Compared with risky driving statuses, such as fatigue or drunk driving, distracted driving has a shorter duration, a more unstable time-varying characteristic, and is more susceptible to the influence of objective conditions [8]. Existing studies have explored distracted driving statuses under the influence of different factors. Cognitive

driving distraction occurs when the driver's attention is subjected to some mental burden or when the driver's attention is on something else, such as carrying on a conversation with a passenger and getting caught up in memories or thoughts [9]. Visual driving distractions can be triggered by anything that takes the driver's focus away from the driving direction; it can be a navigation screen, a mobile phone, or a street sign [10]. Operational driving distractions refer to the driver performing operations unrelated to driving while controlling the vehicle. Typical actions include adjusting onboard equipment, eating, and editing messages [11]. In a real driving process, distracted driving status is rarely of just one form, but a combination of several types [12]. Existing distracted driving status recognition (DDSR) research is mainly carried out based on data reflecting driver behaviour or vehicle motion states.

Methods based on vision-modal data of driver behaviour. With the development of deep learning algorithms, computer vision technology has become popular because of its high accuracy [13]. Additionally, since visual modality data are collected through a camera, there is little interference with the driver while driving; thus, most of the work is carried out based on this scenario. Craye et al. [14] used machine learning methods such as AdaBoost and the hidden Markov model to judge eye, hand, and head movements while driving based on RGB-D map data captured by an active sensor Kinect. In [15], a novel multi-stream long short-term memory (M-LSTM) network was presented by Behera et al. for recognizing driver activities and transforming action information into semantic information. In other works, Eraqi et al. [16] used a genetically weighted ensemble of convolutional neural networks to obtain distraction detection confidence while combining facial and hand motion segmentation. Xing et al. [17] extracted the driver's body from the input background using the Gaussian mixture model (GMM) to recognize actions such as checking the rearview mirror, adjusting a device, and using a mobile phone. Although distracted driving behaviour in some studies was divided into multiple action categories, it essentially includes only two categories: distracted and non-distracted. Such work based on only single-modal visual data is largely limited by its data quality. For example, an occluded driver's body or altered light may reduce accuracy and robustness. In addition, statuses such as cognitive distraction that are not accompanied by significant changes in behavioural actions may be difficult to identify if relying on only single-modal visual data. These behaviours may even be incorrectly identified as a safe driving state, which may create potential safety hazards.

Methods based on sensor-modal data of vehicle motion. Driver behaviour directly affects how the vehicle performs, which can be analyzed by using vehicle motion sensor single-modal data such as lane departure [18], steering wheel angle [19], and longitudinal and lateral acceleration [20]. Such modality data are easy to obtain and have low computational complexity. Lansdown et al. [21] found that multitasking affects the driver's ability to control the vehicle, manifested as an increase in the number of emergency braking events and changes in the steering wheel angle. Additionally, the vehicle state, such as the vehicle's driving trajectory, running speed, and acceleration, will also change with the difficulty that the driver experiences in performing the task [22]. However, methods that rely on only single-modal data from vehicle motion sensing may lead to misjudgments due to noisy data. Moreover, driving style specificity may also affect its recognition accuracy, so adjusting the discrimination threshold appropriately may be necessary for drivers with different driving styles, which will greatly influence the generalization performance.

Methods based on multimodal data. Along with the sustainable development of multi-sensor collection techniques, many studies have been conducted with the fusion of multi-sensor data, such as longitudinal vehicle speed estimation and vehicle localization based on global positioning systems (GPS), inertial measurement units (IMU) and wheel speed sensor (WSS) [23,24], improved autonomous vehicle perception by fusing sensor data from camera and lidar [25], and autonomous emergency braking systems (AEBS) by using lidar, radar, and vision sensors [26]. There have been solutions proposed to integrate multimodal data for identifying distracted driving status, but such works are relatively

limited [27]. Du et al. [28] confirmed that combining facial expression, speech, and car signals provided a better predictive performance for distraction detection. Rashwan et al. [29] proposed a two-stage model. First, three independent modules were used to process signals to extract features from the audio, image, video, and other signals. Then, the driver's dangerous state estimation based on the hidden Markov model was output. Finally, the output and context information of each module were fused through the Bayesian network. Streiffer et al. [30] developed a unified data collection and analysis framework, DarNet, which analyzed driving image data through convolutional neural networks and IMU sensor data through recurrent neural networks. Finally, the two outputs were combined through Bayesian networks. The above multimodal learning approaches utilize multimodal information, in which different modalities can complement each other to improve the recognition system performance, but several challenges remain. It is logical to use appropriate subnetworks for feature extraction for different modality data to extract the spatial information of vision-modal data and the temporal information of motion sensor-modal data. However, it will significantly increase the complexity of later integration. Some early fusion strategies involve simply concatenating multimodal features at the input level, while late fusion strategies perform decision voting. They cannot consider the spatiotemporal dependency for multimodal features; in other words, these fusion approaches are difficult to sustainably learn both intramodal and intermodal correlations.

To solve these issues, a novel multimodal fusion method called vision-sensor fusion transformer (V-SFT) was introduced for recognizing distracted driving statuses with different risk levels, which simultaneously processes the vision-modal of driver behaviour and sensor-modal data of vehicle motion. On the one hand, it may help to reduce the number of accidents and road congestion, thereby reducing traffic pollution and carbon emissions; on the other hand, it may reduce the economic losses caused by traffic accidents, thus saving funds for society and businesses, and has a positive impact on driver assistance technology, as well as economic and environmental sustainability.

The distracted driving risk levels were quantified in this study, including safe driving (no risk), slight risk, moderate risk, and severe risk. This method reflects the driver's safe driving sustainability through the driving risk levels and then provides a basis for allocating or switching the control rights of the human-machine codriving system. The proposed method is composed of three main modules: vision-modal and sensor-modal data early fusion, modality information interaction in the encoder block, and a classifier head for risk level inference. Specifically, feature prefusion was employed to aggregate representations from multiple modality tokens, and then the position-encoded multimodal feature set continuously interact based on the attention mechanism. Finally, the token at a specific position was separated for status classification. During the data acquisition phase, multimodal data during the simulated driving process were recorded synchronously. During the training phase, the model was trained to explicitly describe different risk statuses in distracted driving. To summarize, the main contributions of this study are threefold: (1) The developed end-to-end structure can not only adapt to vision and sensor data simultaneously but can expand and fuse more modal data, (2) feature-level prefusion reduces the complexity of postprocessing operations and provides a prerequisite for human–vehicle information interaction, and (3) taking the risk level as the recognition result to evaluate the status of distracted driving lays the foundation for further research on drivers' driving abilities and driving right allocation or switching.

2. Materials and Methods

2.1. Data Collection

Existing datasets for distracted driving generally come from real vehicles or simulators. For example, the American University in Cairo (AUC) Distracted Driver [31] and State Farm Distracted Driver datasets [32] contain image data of drivers and were collected through offline motion simulation when the vehicle was stationary. The University of Alcalá (UAH)-DriveSet [33] dataset contains sensing data of a vehicle, which were recorded by the inertial

measurement unit (IMU) of an onboard smartphone while the driver imitated a specific driving status in real vehicle experiments. The data recorded by simulator experiments in some other datasets also include eye movement and electrocardiogram (ECG) data, etc., but the intrusiveness of wearable instruments may cause some interference with driving behavior, and their cost is much higher than that of ordinary sensors. Therefore, a new multimodal distracted driving dataset was collected for our research through a rationally designed experiment, which provides both the vision-modal of driver behaviour and sensor-modal data of vehicle motion.

2.1.1. Driving Simulation Platform

Considering the experimental safety, a simulated driving platform was built in the laboratory environment, as shown in Figure 1, which integrates a visual system, a sound system, and a custom I/O board. The driving simulation software UC/win-road was installed on a PC using Microsoft Windows OS. The simulation can simulate driving on various types of roads under various outdoor conditions (weather, light, traffic, etc.) through modelling. The scene was displayed on three LCD screens, and the standard 3D graphics and 120-degree field of view (FOV) provided participants with a realistic driving experience. The interaction between the simulated driving platform and the participants was achieved through the device interface. A Logitech G927 racing steering wheel and pedals were used as input devices. The steering wheel was connected to an active force feedback system, and a passive force feedback mechanism was integrated into the pedals, recreating the feel of the clutch and brake of a real car. In addition, adjustable seats were installed to improve the driving comfort of participants. The data used to infer the driving status were collected by multiple modules, all of which were connected to the host and recorded synchronously. The visual sensor facing the driver's body recorded the RGB video stream of the driver's actions. The "other signals" module equipped with the UC/win-road software output various time-sequential signals such as the positions of the steering wheel, accelerator, brake, and clutch pedals in the form of a LOG file during the experiment. The rest of the hardware included an audio system for voice prompts and an iPad for simulating the central control screen to complete driving tasks. When the simulated driving started, the scene reference objects on the screen and the behaviour of the ego vehicle changed sustainably [34] as the driver manipulated the steering wheel and pedals, and the sound system provided real-time engine roar or brake sounds.

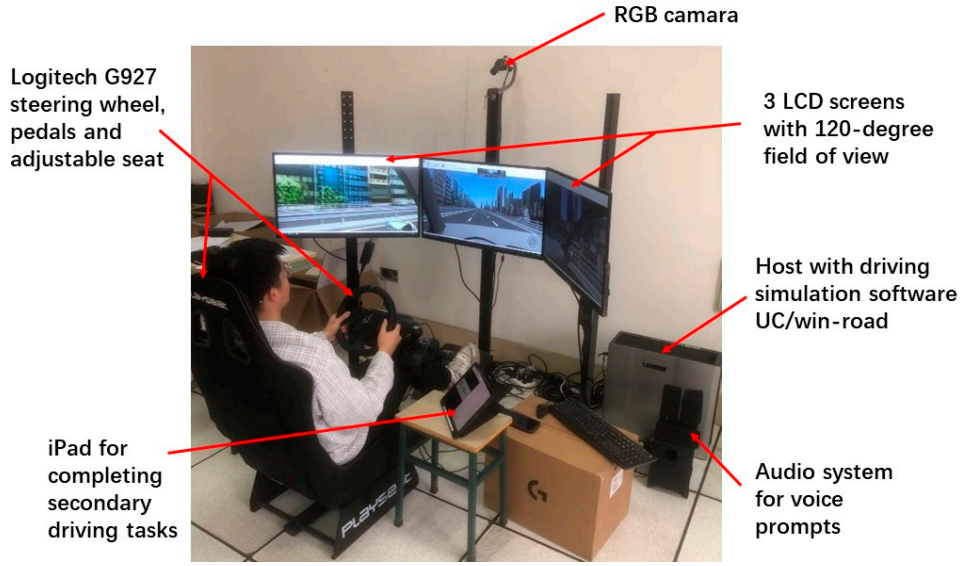
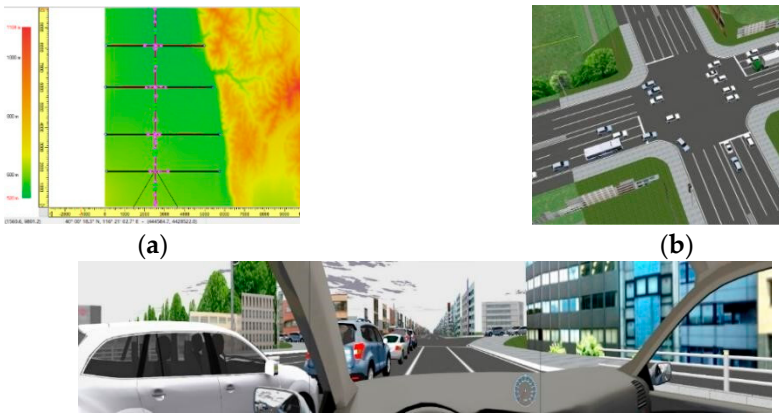


Figure 1. Driving simulation platform. It is composed of display screens, audio system, driving control device, and iPad, and the data of different modalities are collected by RGB camera and equipped software.

2.1.2. Participants and Driving Scenarios

Thirty-six drivers participated in this simulated driving experiment. The average participant age was 25 years old, with a standard deviation of 1.41. Most of the participants were recruited through a social media advertisement, and each participant was paid RMB 50 yuan for their involvement. Both urban driving scenarios and suburban driving scenarios were designed. The two-way three-lane road modelled by asphalt material contained four intersections, the lane width was 3.5 m, and the total distance was 10 km, as shown in Figure 2a,b. Figure 2c shows the driver's perspective at an urban road intersection. Participants were asked to complete simulated driving on each route at an speed of approximately 60 km/h, and each route was driven only once.



(**a**) (**b**)

(**c**)

Figure 2. Driving scenario design: (**a**) layout of the road; (**b**) overlook of the intersection with traffic lights, each road is a two-way three-lane road and a certain traffic flow is set; and (**c**) driver's perspective at an urban road intersection.

2.1.3. Secondary Task

To simulate the contingency and suddenness of the distracted driving status as realistically as possible, when the participant drove the simulated vehicle to a trigger point on the road, a specific location designed in advance but unknown to the participant, a sound prompt was triggered, and the participant needed to complete the secondary task according to the prompt while driving. Extensive work has demonstrated that the great interference is predicted when a secondary task shares the same sensory modality as the primary driving task, since both tasks require access to the same limited pool of resources [35]. Therefore, the limited resources of the primary task are occupied to varying degrees by designing different combinations of secondary tasks to induce a distracted driving status with different risk levels. Four secondary tasks of different complexity were designed, and the corresponding driver behaviours are shown in Figure 3a–d. These behaviours were triggered by Task_0, Task_1, Task_2, and Task_3, respectively. The specific task combination is described in Table 1.

(**a**) (**b**) (**c**) (**d**)

Figure 3. Driver behaviour under different secondary tasks: (**a**) Task_0; (**b**) Task_1; (**c**) Task_2; and (**d**) Task_3.

Table 1. Secondary task descriptions of varying levels of complexity.

Task Category	Task Type	Description
Task_0	No secondary tasks	No voice command.
Task_1	Cognitive distractions	Triggering the voice command, following the command to calculate the two-digit addition operation that reported in the command, and speaking the result.
Task_2	Cognitive distractions & Visual distractions	Triggering the voice command, following the command to observe the two-digit addition operation that displayed on the screen, and speaking the result.
Task_3	Cognitive distractions & Visual distractions & Operating distractions	Triggering the voice command, following the command to observe the two-digit addition operation that displayed on the screen, and inputting the result by handwriting on the screen.

2.1.4. Procedure

When participants arrived at the driving simulator laboratory, they were informed of the purpose of the experiment and related arrangements in detail before signing up to participate. Before the start of the formal experiment, participants performed some practices to adapt to the manipulation of the driving simulation platform, completed the physical condition questionnaire, and confirmed the right to terminate the experiment at any time. In the formal experiment, participants needed only to follow the system prompts and drive to the end along a set route, without the interference of external factors such as experimenters. It took approximately 15 min to complete a route. The physical condition was fed back again while the participant took a break between the two drives. Each participant needed to complete the four simulated driving routes on two road types in total. Even with the same road type, the driving environments of the two routes were different. To avoid predictability and minimize practice effects, the secondary tasks were distributed on the four designed roads in a counterbalanced order [36] by a Latin square design (in Table 2). As well, the driving order of the four roads was also distributed evenly for each participant. Figure 4 illustrates the secondary task sequence on road one and the recorded data of two modalities, the vision-modal data of the driver's behaviour and the sensor-modal data of the vehicle motion.

Table 2. Secondary task arrangement on different routes.

	Trigger Point 1	Trigger Point 2	Trigger Point 3	Trigger Point 4
Road 1 (Urban)	Task_2 (27 + 54)	Task_1 (14 + 17)	Task_3 (49 + 25)	Task_0 (/)
Road 2 (Urban)	Task_1 (33 + 19)	Task_0 (/)	Task_2 (36 + 48)	Task_3 (25 + 16)
Road 3 (Suburban)	Task_3 (28 + 34)	Task_2 (37 + 19)	Task_0 (/)	Task_1 (18 + 36)
Road 4 (Suburban)	Task_0 (/)	Task_3 (43 + 18)	Task_1 (27 + 15)	Task_2 (17 + 26)

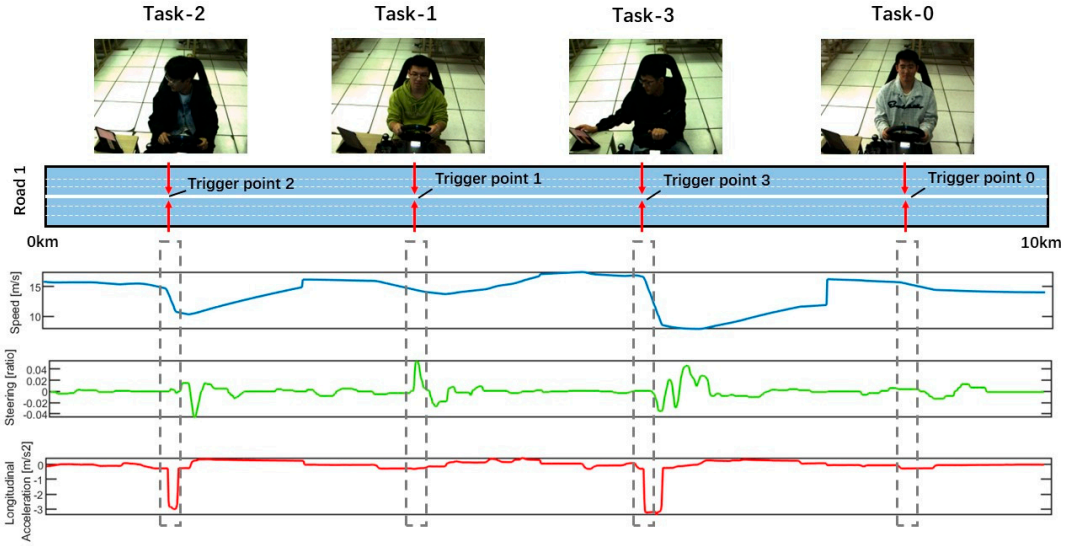


Figure 4. Secondary task sequence and data recording on road one.

2.2. Methods

The raw data collected in the previous section were processed by data cleaning, extraction, merging, and standard normalization, to obtain a distracted driving status dataset with risk levels. This dataset contained two data modalities. The visual-modal data were acquired by image extraction from the video stream of the driver's behaviour recorded by the camera. The sensor-modal data expressing the vehicle's motion state were synchronously recorded through the simulated driving software. There were failure scenarios for status recognition based on only visual data, such as thinking, listening to the radio, and other cognitively distracted driving statuses, because these driving statuses usually do not have apparent changes in visual features. At the same time, relying on only the vehicle motion sensor modality is too sensitive to signal fluctuations, lacks information interaction with the driver, and has insufficient robustness and accuracy. Therefore, an attempt was made to fuse the two data modalities to obtain a joint representation and use the complementary human–vehicle information to enhance the discriminative feature learning. In our previous research [37], superior results were achieved in the distracted driving action classification based on a vision transformer (ViT) [38]. Inspired by this, a vision-sensor fusion transformer (V-SFT) fusion strategy for early fusion and information interaction of multimodal data is proposed. Figure 5 shows the overall architecture of the model. The model input is the preprocessed human–vehicle multimodal data pair, and the output is the risk level of the distracted driving status. The model can be described as three modules.

2.2.1. Module 1: Early Fusion of Vision-Modal Data and Sensor-Modal Data

First, the vision-modal data of driver behaviour were tokenized, as shown in Figure 6. The specific steps are as follows. (1) Resizing the raw image to [224 × 224], converting it into a tensor, and then normalizing its mean and variance to 0.5. (2) Dividing the [3 × 224 × 224] RGB image into 196 [3 × 16 × 16] patches using a convolution kernel with a size of 16, a stride of 16, and 768 convolution kernels. (3) All the patches are flattened and then transposed into a one-dimensional sequence, that is, 196 one-dimensional image tokens with a length of 768. The following Formula (1) describes the process:

$$\left[X_1^i,\ X_2^i,\ X_3^i,\ \ldots, X_{196}^i\right] = T\{Flatten[Conv2D(Normalize(X_{image}),\ k = (16,16),\ s = 16)]\} \quad (1)$$

where X_{image} represents the input visual modality data with a shape of [3 × 224 × 224]; k and s represent kernel size and stride, respectively; $Conv2D$ denotes the two-dimensional

convolution; T indicates transpose; and X^i represents a one-dimensional token with a shape of [1 × 768], for a total of 196.

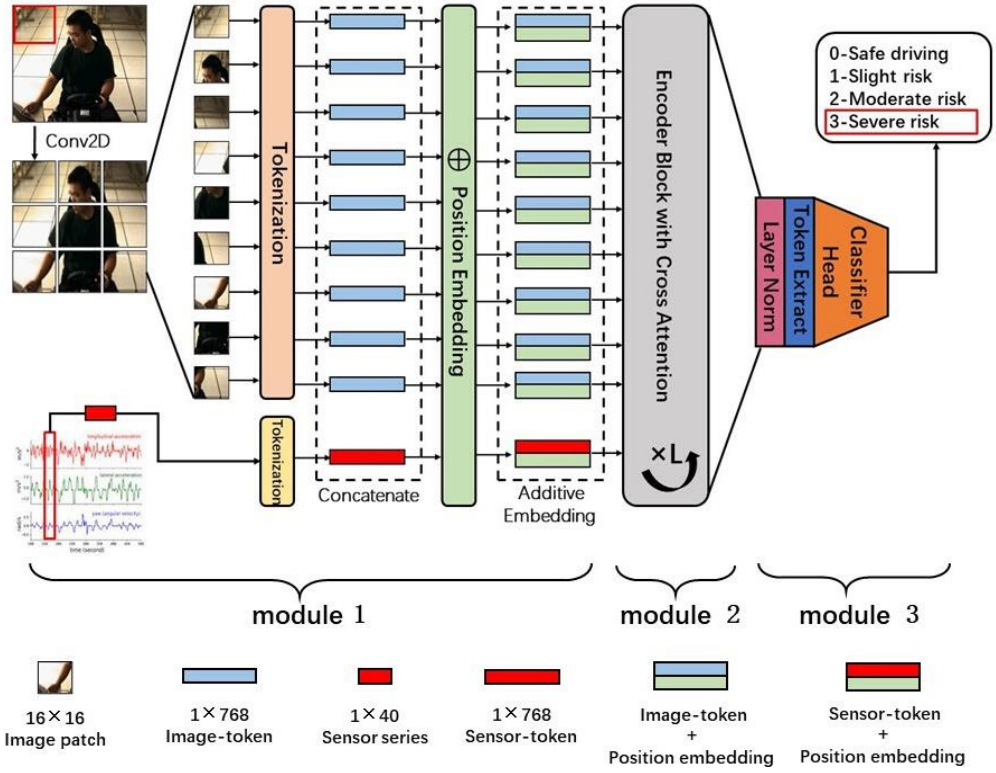


Figure 5. Framework of the V-SFT model for risk level recognition.

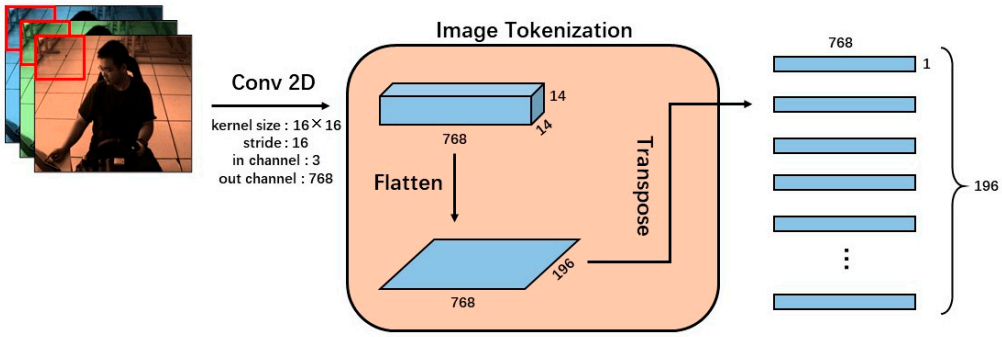


Figure 6. Tokenization of vision-modal data.

The second step is the tokenization of vehicle motion sensor-modal data, as shown in Figure 7. The sensor-modal data in the input data pair were standardized during preprocessing. Although its initial shape is a one-dimensional series of [1 × 40], its feature size is far from the vision-modal token with a total size of [196 × 768]. If the two are directly embedded, the vision-modal token with a larger number of features may weaken the information expression of the sensor-modal token or even completely cover it. Therefore, a trainable self-learning vector w-token is introduced with a size of [40 × 768], and Xavier initialization [39] is performed to upsample the [1 × 40] sensor series and project it to a [1 × 768] vector space. The obtained vector with shape [1 × 768] can be regarded as the class-token in the vision transformer (ViT) model. However, the improvement is that, compared with the class-token, which does not contain any initial information in the original ViT model, the [1 × 768] obtained vector here contains the prior information of the vehicle's motion characteristics, which could be the key to improving model performance. In addition, to preserve the advantages of pretraining on large datasets, the pretrained class-

token of the original ViT model is also embedded to obtain the final [1 × 768] sensor token. The token contains both the vehicle motion information and the pretrained parameters, which can improve the model performance and speed up model fitting accordingly. The following Formula (2) describes the process:

$$X_{cla}^{s} = (X_{sensor} \otimes W_{token}) \oplus Cla_{token}^{pre} \tag{2}$$

where X_{sensor} represents the [1 × 40] input sensor-modality series, W_{token} represents the [40 × 768] trainable self-learning vector, Cla_{token}^{pre} represents the [1 × 768] pretrained class-token, X_{cla}^{s} represents the [1 × 768] sensor token, ⊗ represents matrix multiplication, and ⊕ represents the matrix addition.

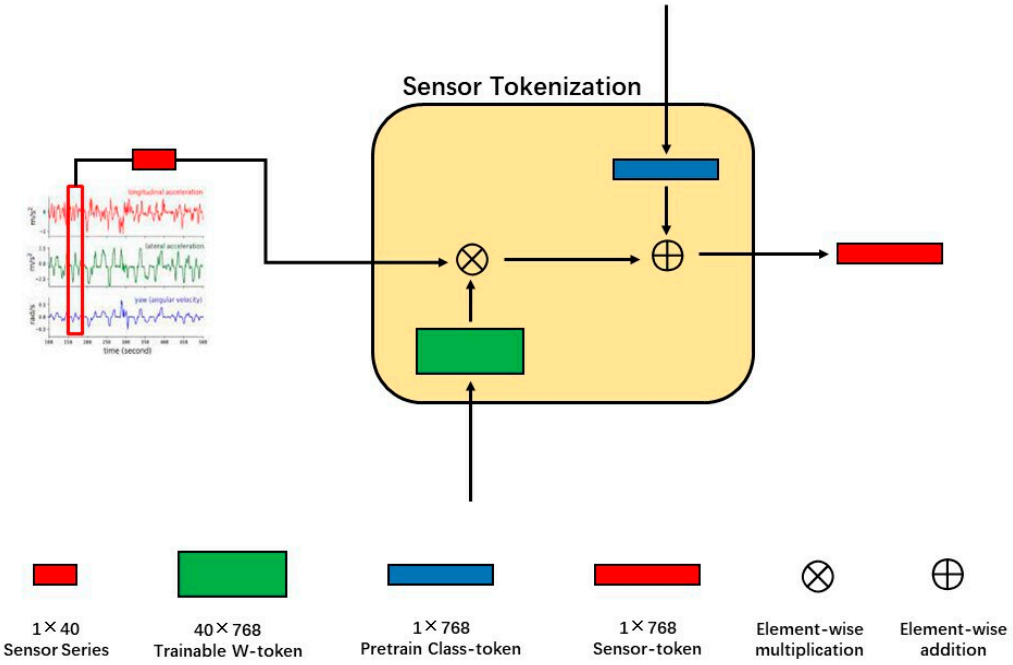


Figure 7. Tokenization of sensor-modal data.

The third step is to obtain the joint representation of the two modalities. The acquired [196 × 768] image token and [1 × 768] sensor token are concatenated to output a [197 × 768] sequence group, that is, 197 sequences of shape [1 × 768]. However, at this time, these sequences are not position encoded. The position information is embedded by adding the trainable parameter position embedding to the [197 × 768] sequence group element by element. Finally, the joint token output from two modal data fusion is obtained. The above process can be summarized as Formulas (3) and (4):

$$\left\{ X_{cla}^{s}, \left[X_1^i, X_2^i, X_3^i, \ldots, X_{196}^i \right] \right\} = Cat\left(X_{token}^{sesor}, X_{token}^{image} \right) \tag{3}$$

$$X_{token}^{joint} = \left\{ X_{cla}^{s}, \left[X_1^i, X_2^i, X_3^i, \ldots, X_{196}^i \right] \right\} \oplus PE \tag{4}$$

where Cat denotes vector concatenation, $\{ X_{cla}^{s}, [X_1^i, X_2^i, X_3^i, \ldots, X_{196}^i] \}$ represents a sequence group, PE denotes trainable vector position embedding with the same size, X_{token}^{joint} represents the joint token of the two modal tokens, and ⊕ denotes the matrix addition.

2.2.2. Module 2: Modality Information Interaction in the Encoder Block

As shown in Figure 8, the joint representation X_{token}^{joint} enters the repeatedly stacked encoder blocks after passing through a dropout [40] layer. Each encoder block contains a residual multihead self-attention (MSA) block and a residual multilayer perceptron (MLP) block. Specifically, layer norm (LN) is applied before every MSA and MLP, and then a

dropout layer and a residual connection [41] are applied after them in sequence. In this module, the sensor-modal token X^s_{cla} representing vehicle motion information and the vision-modal token $\left[X^i_1,\ X^i_2,\ X^i_3,\ \ldots, X^i_{196}\right]$ representing driver behaviour information, and these tokens interact continuously to learn effective abstract representations.

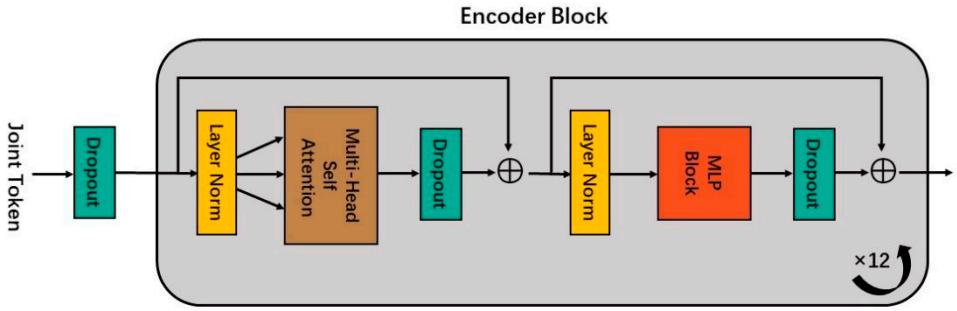


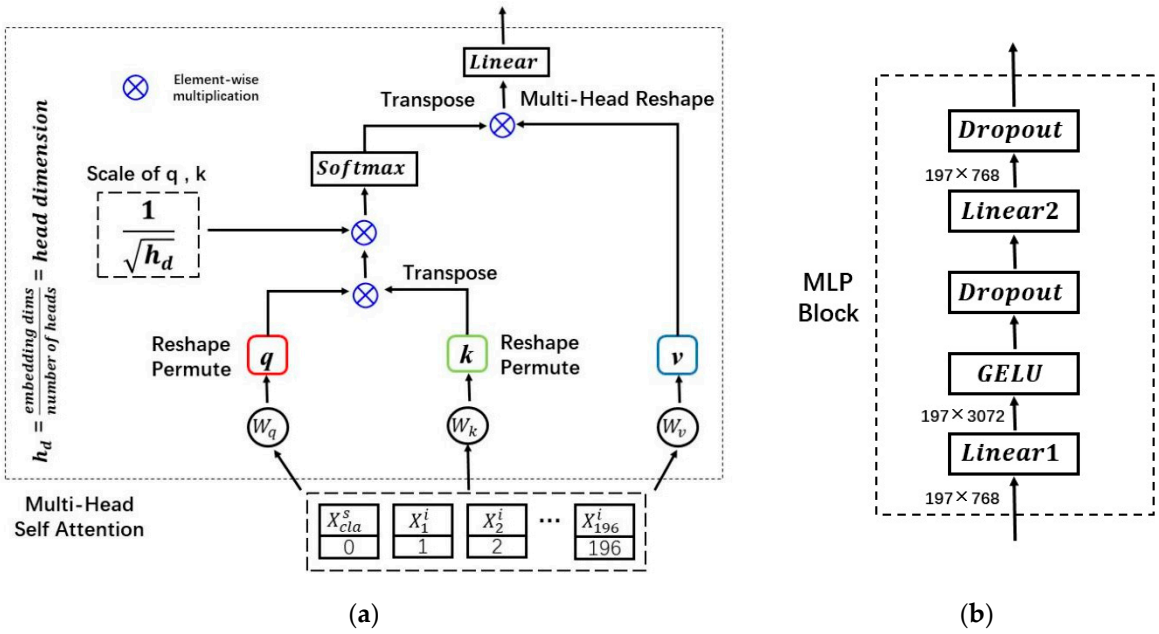
Figure 8. The internal structure of the encoder block.

The interaction of different modal information in the MSA block is shown in Figure 9a. The following Formula describes the process:

$$q^i = DP\left(X^{joint}_{token}\right)W_q = \frac{Linear\left[DP\left(X^{joint}_{token}\right)\right]_{clip[0]}}{number\ of\ heads} \tag{5}$$

$$k^i = DP\left(X^{joint}_{token}\right)W_k = \frac{Linear\left[DP\left(X^{joint}_{token}\right)\right]_{clip[1]}}{number\ of\ heads} \tag{6}$$

$$v^i = DP\left(X^{joint}_{token}\right)W_v = \frac{Linear\left[DP\left(X^{joint}_{token}\right)\right]_{clip[2]}}{number\ of\ heads} \tag{7}$$



(**a**) (**b**)

Figure 9. (**a**) Multimodal information interaction based on multi-head self-attention and (**b**) MLP block.

In Formulas (5)–(7), q^i, k^i, and v^i are the query, key, and value, respectively, corresponding to the i-th head (branch) of the MSA block, all of which are [197 × 96] vectors. They can be obtained by mapping [197 × 768] X^{joint}_{token} through [768 × 96] vectors W_q, W_k, and W_v. The

specific steps are as follows. First, the input [197 × 768] X_{token}^{joint} is linearly transformed into [197 × 768 ×3] through the fully connected layer. Then, through reshaping, permutation, and clipping in turn, [197 × 768 × 3] is divided into three [197 × 768] vectors according to the indexes [0], [1], and [2]. Finally, each [197 × 768] vector is equally divided according to the number of heads; the number of heads here is eight.

$$head_i = Attention\left(q^i, k^i, v^i\right) = Softmax\left(\frac{q^i\left(k^i\right)^T}{\sqrt{h_d}}\right)v^i \tag{8}$$

In Formula (8), q^i, k^i, and v^i interact with each other based on the attention mechanism [42] to obtain the output of each [197 × 96] branch $head_i$, where $\left(k^i\right)^T$ is the transpose of k^i and h_d is the dimension of each head, which is 96. $Softmax$ denotes the softmax functions.

$$MSA = [Concat(head_1, head_2, \ldots, head_i)]W_O \tag{9}$$

$$Res_{MSA} = DP\left(X_{token}^{joint}\right) \oplus DP(MSA) \tag{10}$$

In Formula (9), after concatenating the output of eight heads, the sequence group $[head_1, head_2, \ldots, head_8]$ is [197 × 768]. Then, a [768 × 768] vector W_O is used for linear projection to obtain MSA, which is still [197 × 768]. Last, after MSA passes through a dropout layer, it is added to the $DP\left(X_{token}^{joint}\right)$ to obtain the Res_{MSA} with a shape of [197 × 768] in Formula (10), which is the output of the entire residual MSA block. The MLP block contains two linear layers with Gaussian Error Linear Unit (GELU), as shown in Figure 9b. First, Linear1 transforms the input [197 × 768] sequence Res_{MSA} into [197 × 3072]. Then, after the GELU activation function and dropout, the [197 × 3072] sequence is retransformed back to [197 × 768] by Linear2 and then passes through a dropout layer. Similarly, the output of the residual MLP block is still [197 × 768].

2.2.3. Module 3: Classifier Head for Risk Level Inference

The structure of module 3 is shown in Figure 10. A norm layer is also applied before the classifier head. From the previous section, after 12 consecutive encoder blocks, the output shape is still [197 × 768]. However, at this time, the spatiotemporal dependency of both intramodality and intermodality has been learned. Then, extracting the category sequence class-token in this sequence group is necessary [38]. Since position information encoding was performed in advance in module 1, the [1 × 768] sequence whose index is [0] is removed. Finally, the [1 × 768] features are mapped into four label spaces of risk level by a linear layer, and the number of neurons in the output layer is set to four. The class probability distribution is normalized using the softmax function, as shown in Formula (11).

$$softmax\left(z_j\right) = \frac{e^{z_j}}{\sum_{c=1}^{C} e^{z_c}} \tag{11}$$

where $softmax\left(z_j\right)$ represents the value of j after the C-dimension vector z is mapped by the softmax functions. C represents the number of classes.

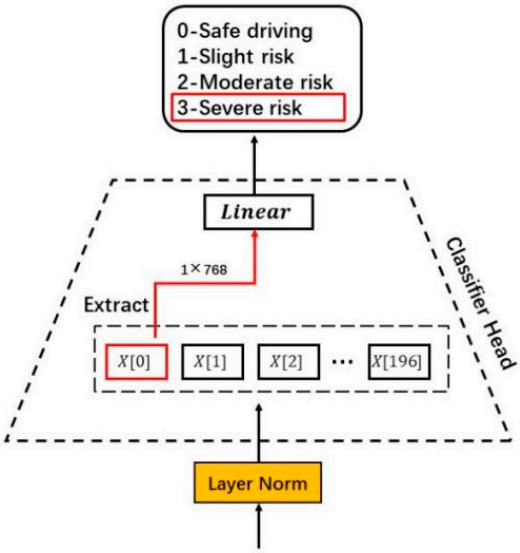


Figure 10. Structure of the classifier head for risk level inference.

2.3. Evaluation Metrics

To quantitatively measure the testing results, precision, recall, and F1 score were applied to measure the model performance [43]. Accuracy is the overall prediction accuracy of the model. Specifically, four classification results are defined and statistically calculated for each behaviour, including true positive (TP_i, is samples of behaviour i that are correctly identified), true negative (TN_i, denotes that samples that do not belong to behaviour i are classified into other behaviours), false positive (FP_i, is cases in other behaviours that are incorrectly classified into behaviour i), and false negative (FN_i, is cases in behaviour i that are incorrectly predicted as other behaviours). Specifically, the precision of behaviour i (P_{ri}, $i = 1, 2, 3 \ldots$) is calculated as:

$$P_{ri} = \frac{TP_i}{TP_i + FP_i} \times 100\% \tag{12}$$

The recall of behaviour i (R_{ei}, $i = 1, 2, 3 \ldots$) is calculated as:

$$R_{ei} = \frac{TP_i}{TP_i + FN_i} \times 100\% \tag{13}$$

The F1-score ($F1$) depends on both P_{ri} and R_{ei} and is the harmonic mean of these two values.

$$F1 = 2 \times \frac{P_{ri} \times R_{ei}}{P_{ri} + R_{ei}} \times 100\% \tag{14}$$

Finally, the accuracy (Acc) is calculated as:

$$Acc = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \times 100\% \tag{15}$$

3. Results

In this section, the proposed V-SFT model for distracted driving status risk level recognition was trained and tested. First, in the case of relying on only vision single-modal data, the evaluation results of the proposed V-SFT and some convolutional neural network (CNN) models on the collected dataset were compared. Next, in the case of relying on only sensor single-modal data, the evaluation results of the V-SFT and some benchmark models on the collected dataset were compared. Last, in the case of using both modalities simultaneously, the evaluation results of the proposed V-SFT fusion model on the collected

dataset were compared with some other fusion structures. After preprocessing the raw data of the two vision-sensor modalities in the dataset, the training, validation, and test sets were divided at a ratio of 7:2:1, and the corresponding ratio of the sample numbers was 16,800:4800:2400. In order to avoid the influence of other factors, the implementation details, such as hardware platform, training strategy, and hyperparameters, were all set to be consistent during the training process. Specifically, the training and testing of all models were supported by GPU NVIDIA GeForce RTX 3080 Ti. In terms of training strategies and hyperparameter settings, pretraining was performed on the mixed distracted driving datasets of AUC [31] and State Farm [32], and then training was continued for 100 epochs on the collected dataset based on transfer learning. In addition, the batch size was set to 32, and the learning rate was set to 0.001. During the training process, a large amount of labeled data were fed into the model, and the four risk level labels of safe driving, slight risk, moderate risk, and severely risk were used as fitting targets, which were encoded as [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1].

3.1. Experimental Evaluation of Vision-Only Modality Input

In the case of inputting only vision single-modal data, the validation curves during the training process of the proposed V-SFT (v-only) and several CNN benchmark models, such as MobileNetV3 [44], InceptionV3 [45], and ResNet34 [41] are shown in Figure 11. When the trend of the curve is relatively stable, the validation accuracy of the V-SFT with only vision-modal input (V-SFT v-only) is slightly higher than that of ResNet34 and much higher than that of MobileNetV3 and InceptionV3. This may be largely attributed to the ability of the V-SFT backbone to capture long-range dependencies rather than local features [46]. As shown in Table 3, the results on the test set, specifically precision (*Pr*), recall (*Re*), F1-score (*F*1) and accuracy (*Acc*), were calculated based on the evaluation metrics. They are slightly lower compared to the validation process but still within the expected range.

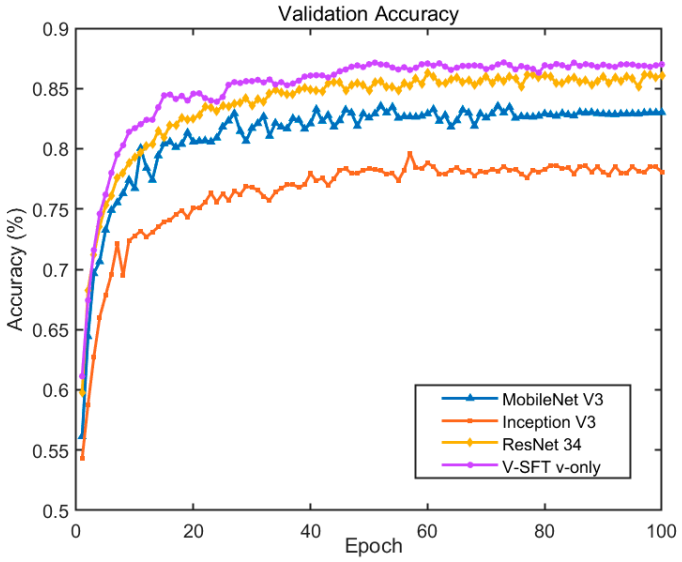


Figure 11. Validation accuracy curves of different models with only vision-modality input.

The confusion matrixs of the test results are shown in Figure 12. The recognition performance of the four models with single-modal visual data input can be intuitively observed through the heatmap. It was shown that all the models were able to identify moderate and severe risk statuses well relying on only vision-modal data. The moderate risk status triggered by task_2 includes not only the cognitive distraction caused by thinking about calculations but also the visual distraction caused by observing the screen through gaze or head movement. In addition to cognitive and visual distractions, the severe risk status triggered by task_3, also includes operational distractions added by manual screen manipulation. It can be seen that all models capture the differences in actions well and

make distinctions. The best performance is achieved by the proposed V-SFT (v-only), which has a test accuracy rate of 86.5% for the four statuses. Compared with MobileNetV3, InceptionV3, and ResNet34, it is 3.8%, 9.2%, and 0.8% higher, respectively.

Table 3. The test results of different models with only vision-modal input.

Models	Safe Driving			Slight Risk			Moderate Risk			Severe Risk			*Acc* (%)
	Pr (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	
MobileNet V3	**77.1**	78.7	77.9	76.3	**76.7**	**76.5**	87.4	85.8	86.6	90.3	89.7	90.0	82.7
Inception V3	66.6	74.5	70.3	69.6	59.5	64.2	89.7	84.3	86.9	83.7	91.0	87.2	77.3
ResNet 34	75.0	78.8	76.9	76.7	73.2	74.9	96.2	93.0	94.6	95.0	97.7	96.3	85.7
V-SFT v-only	76.2	**80.0**	**78.1**	**77.8**	74.3	76.0	**96.7**	**93.3**	**95.0**	**95.3**	**98.2**	**96.7**	**86.5**

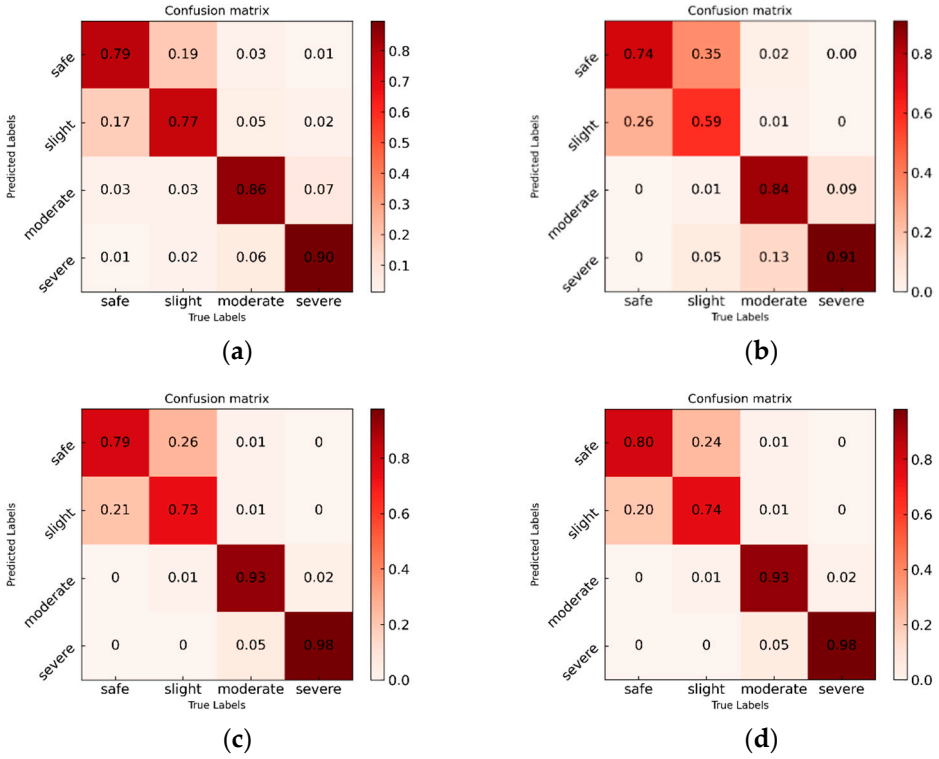


Figure 12. The confusion matrixs of different models with only vision-modal input: (**a**) MobileNetV3; (**b**) InceptionV3; (**c**) ResNet34; (**d**) V-SFT v-only.

It can be seen that the above four models that rely on only vision-modal data may confuse some safe driving statuses with slight risk statuses, and even the best-performing V-SFT (v-only) has not overcome this problem. Specifically, when there was no secondary driving task (task_0), the driver continued to hold the steering wheel with both hands and look straight ahead. In the slight risk status triggered by task_1, the driver only opened his mouth slightly when answering the mathematical calculation given by the voice prompt. Therefore, it is speculated that when relying on only the vision modality, the model relies heavily on illumination or pixel definition in learning basic features such as lines, edges, and colors. Particularly under real driving conditions, it would be unimaginable if a large number of slight risk statuses were misjudged as safe driving, which is also an issue that this study will solve.

3.2. Experimental Evaluation of Sensor-Only Modality Input

In the case of inputting only sensor single-modal data, the validation curves during the training process of the proposed V-SFT (s-only) and several benchmark models, such as support vector machine (SVM) [47], random forest [48], recurrent neural network (RNN) [49], gated recurrent unit network (GRU) [50], and long short term memory network (LSTM) [51], are shown in Figure 13. It can also be seen that after the curve fluctuation is relatively stable, the validation accuracy of the V-SFT with only sensor-modal input (V-SFT s-only) is higher than the others. As shown in Table 4, the test results of the above six models on the collected dataset, specifically precision (*Pr*), recall (*Re*), F1-score (*F*1) and accuracy (*Acc*), were calculated based on the evaluation metrics in the previous section. Among them, the V-SFT model with only sensor-modal input (V-SFT s-only) achieves the highest test accuracy of 75.0%. Compared with SVM, random forest, RNN, GRU, and LSTM, the test accuracy of V-SFT (s-only) is 5.2%, 6.0%, 3.5%, 2.5%, and 2.9% higher, respectively. This also benefits from the transformer backbone of the V-SFT, which takes advantage of the temporal representation of sensor-modal data [52].

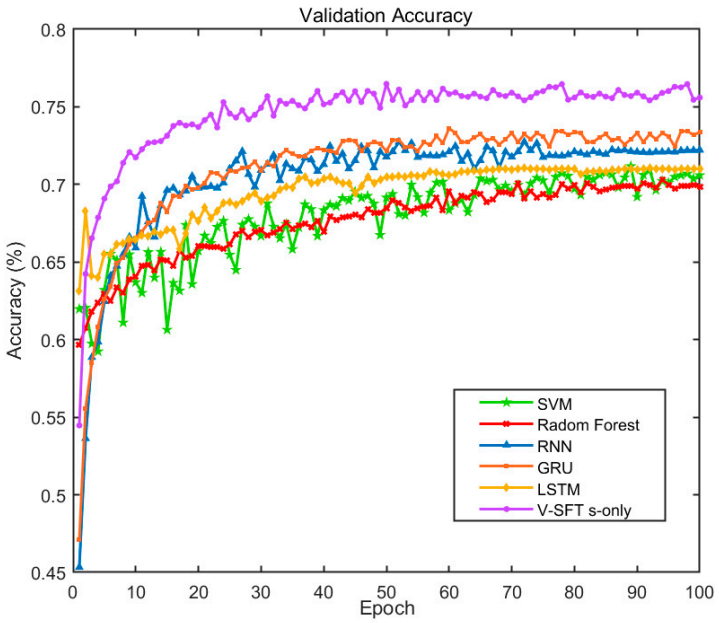


Figure 13. Validation accuracy curves of different models with only sensor-modality input.

Table 4. The test results of different models with only sensor-modal input.

Models	Safe Driving			Slight Risk			Moderate Risk			Severe Risk			*Acc* (%)
	Pr (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	
SVM	69.3	72.8	71.0	72.2	71.5	71.8	69.8	62.0	65.7	67.9	72.7	70.2	69.8
Radom Forest	64.5	67.8	66.1	72.4	71.7	72.0	73.0	64.8	68.7	67.1	71.8	69.4	69.0
RNN	77.4	69.8	**73.4**	**73.9**	**84.7**	**78.9**	66.6	63.5	65.0	68.0	68.0	68.0	71.5
GRU	59.5	**81.0**	68.6	72.0	51.5	60.0	**83.6**	73.0	**77.9**	**80.3**	84.3	**82.3**	72.5
LSTM	60.3	79.8	68.7	72.3	52.7	61.0	81.2	70.7	75.6	79.0	**85.2**	82.0	72.1
V-SFT s-only	**78.1**	69.0	73.3	73.3	82.2	77.5	76.8	**74.7**	75.7	72.5	74.2	73.3	**75.0**

The confusion matrixs of the six models with sensor single-modal data input are shown in Figure 14. Compared to the model that relies on only the vision modality in the previous section, the test accuracy that relies on only the sensor modality is much lower. The reason may be that the feature dimension of sensor data is relatively smaller than that of vision data. In addition, vehicle motion sensor data are easily affected by factors such

as road conditions and driving styles, so the vehicle information that can be learned may not be sufficient. Its direct manifestation is that some moderate risk statuses (triggered by task_2) are confused with severe risk statuses (triggered by task_3) when relying on only sensor-modal data. It is speculated that the feature difference of the vehicle motion sensor modality is not as significant as that of the visual modality in these two levels. However, for the distinction between safe driving (triggered by task_0) and slight risk (triggered by task_1), the V-SFT model with only sensor-modal input (V-SFT s-only) showed some improvement. The model is also superior to the other five models in overall recognition accuracy. Therefore, utilizing vehicle information in the sensor modality to supplement driver information in the vision modality also serves as motivation for this study.

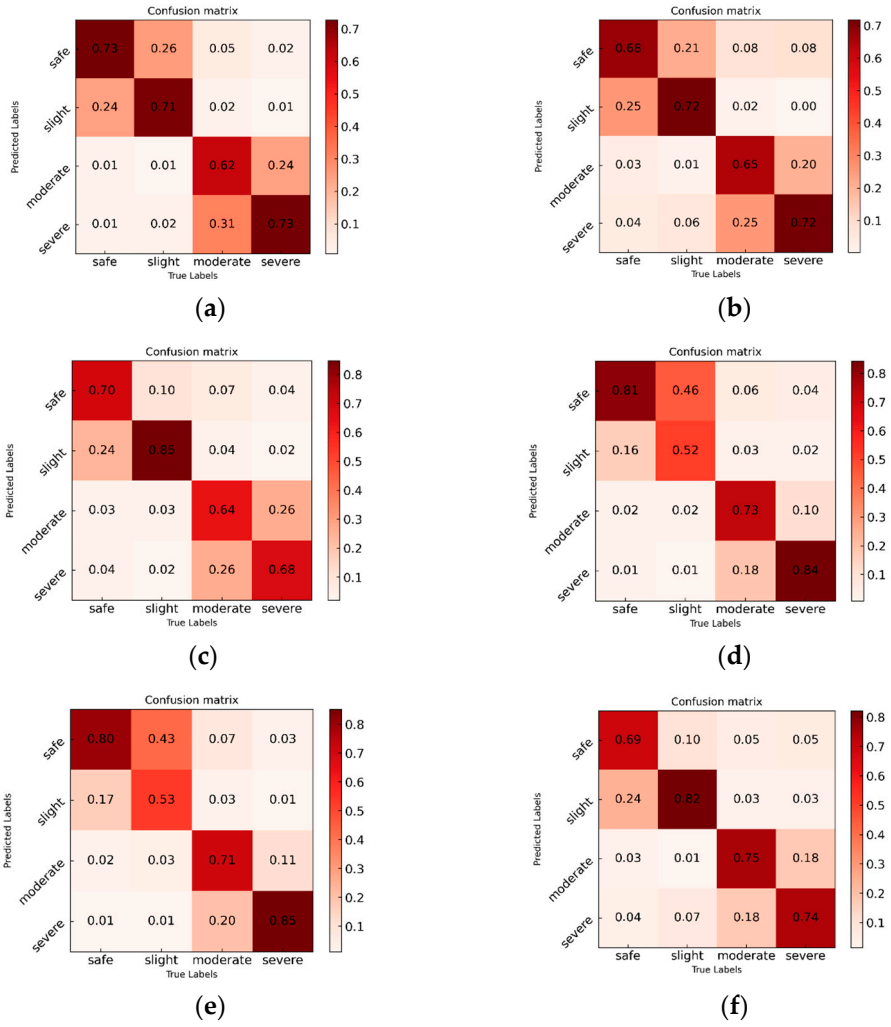


Figure 14. The confusion matrixes of different models with only sensor-modal input: (**a**) SVM; (**b**) Radom Forest; (**c**) RNN; (**d**) GRU; (**e**) LSTM; and (**f**) V-SFT s-only.

3.3. Experimental Evaluation for Vision-Sensor Multimodal Data Input

In this section, under the premise of simultaneous input of vision-sensor multimodal data, CNN-RNN, CNN-GRU, and CNN-LSTM fusion models were built based on the early fusion strategy and compared with the proposed V-SFT on the collected dataset using simultaneous vision-sensor multimodal data input. Figure 15 shows the validation accuracy curves of the above models during the training process; it can be seen that the curve of the V-SFT model is higher than that of the other models after stabilization. The test results, specifically precision (*Pr*), recall (*Re*), F1-score (*F*1) and accuracy (*Acc*), are also calculated based on the evaluation matrics, as shown in Table 5. In addition, the parameters

of the four fusion models CNN-RNN, CNN-GRU, CNN-LSTM and V-SFT are 81.40 M, 77.92 M, 81.26 M, and 86.33 M respectively.

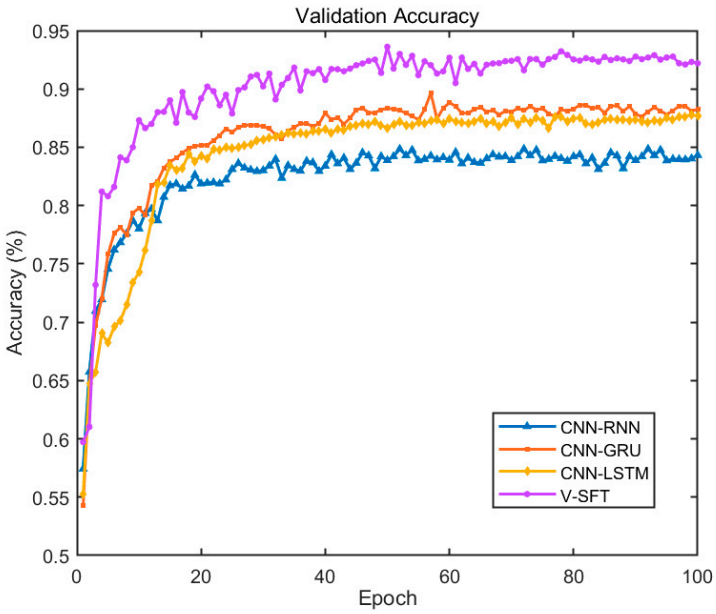


Figure 15. Validation accuracy curves of different models with vision-sensor multimodality input.

Table 5. The test results of different models with vision-sensor multimodal input.

Models	Safe Driving			Slight Risk			Moderate Risk			Severe Risk			*Acc* (%)
	Pr (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	*Pr* (%)	*Re* (%)	*F*1 (%)	
CNN-RNN	76.6	82.5	79.4	80.0	72.8	76.2	85.8	89.7	87.7	92.9	90.0	91.4	83.8
CNN-GRU	86.4	76.0	80.9	77.7	87.5	82.3	93.5	90.8	92.1	92.8	94.8	93.8	87.3
CNN-LSTM	85.1	75.3	79.9	76.8	86.0	81.1	93.7	91.2	92.4	93.1	95.2	94.1	86.9
V-SFT	**90.5**	**83.8**	**87.0**	**83.9**	**90.2**	**86.9**	**97.8**	**94.5**	**96.1**	**96.3**	**99.3**	**97.8**	**92.0**

First, V-SFT achieves the best overall recognition accuracy of 92.0% among all fusion models. Compared with CNN-RNN, CNN-GRU, and CNN-LSTM, the recognition accuracy of V-SFT is 8.2%, 4.7% and 5.1% higher, respectively. Considering that the input of all the above models is vision-sensor multimodal data and they are all built based on the early fusion strategy, it can be inferred that the structural difference is the main reason affecting model performance. The proposed V-SFT uses one unified transformer backbone to simultaneously process two modalities and to preserve the intermodality dependencies. Second, referring to the previous Tables 3 and 4, it can be found that the V-SFT using vision-sensor multimodal input achieves significantly improved results compared to single-modal data input. Specifically, the recognition accuracy of the V-SFT with multimodal fusion input improves by 5.5% compared with that relying on only vision single-modal input; compared with that relying on only sensor single-modal input, it improves by approximately 17%. This indicates that fusing multimodal information is an effective method for improving driving status understanding.

The confusion matrixs of the four fusion models using vision-sensor multimodal data input are shown in Figure 16. Compared to the previous Figure 12, it can be observed that after increasing the sensor-modality input, the recognition effect of the safe driving status and the slight risk status greatly improves. Similarly, referring to the Figure 14, it can be observed that after increasing the vision-modality input, the recognition effect for the moderate risk status and the severe risk status also improves.

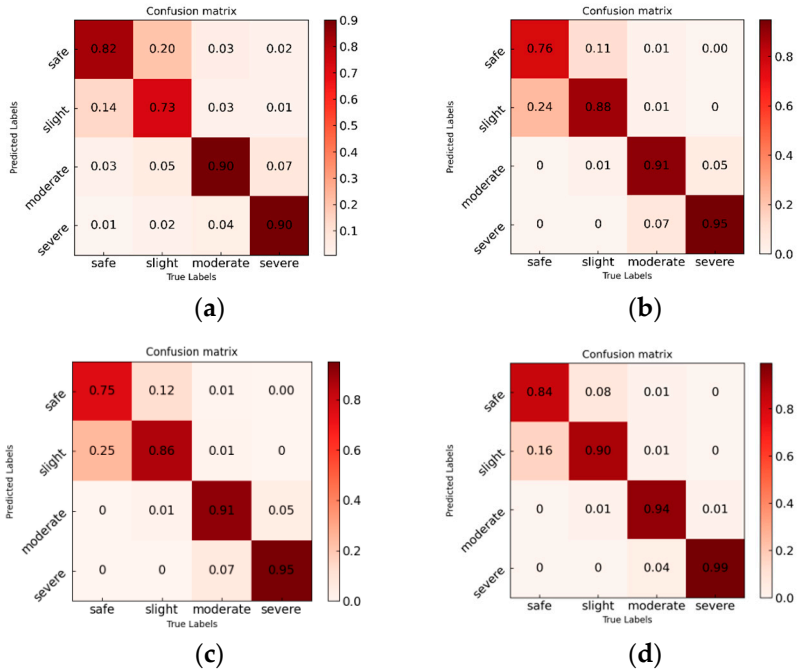


Figure 16. The confusion matrixes of different models with vision-sensor multimodal input: (**a**) CNN-RNN; (**b**) CNN-GRU; (**c**) CNN-LSTM; and (**d**) V-SFT.

4. Discussion

The purpose of this study is to quantify and identify the distracted driving risk levels, including safe driving (no risk), slight risk, moderate risk, and severe risk. By accurately identifying the risks, sustainable driving safety can be achieved, which can effectively improve public health and traffic flow efficiency. In this experiment, driving statuses with different risk levels were triggered by rationally designing secondary tasks. Thirty-six volunteers were recruited to simulate driving on two road types and four routes with different orders of secondary tasks, and data from different modalities were recorded simultaneously. In addition, a V-SFT model was constructed, and the two modal data types were adapted through one unified backbone network. The vision-modal data representing the driver's behaviour and the simultaneously recorded sensor-modal data representing the vehicle's motion interacted and fused.

The performance of the method based on multimodal fusion significantly improved compared with that based on a single modality. Specifically, although some works that use only visual single-modal data for distracted driving status recognition (DDSR) can achieve high accuracy on their datasets [16], there are still some failure scenarios, such as cognitive distraction in real driving conditions [9]. As a frequently occurring risky driving status, cognitive distraction usually presents as a temporary decline in the perception and reasoning ability [22]. In the results of Table 3, it can be seen that for the test results of slight risk status represented by cognitive distraction, the highest recognition recall rate of the methods relying on visual single-modality data, which is 76.7%. The reason may be that the slight risk status represented by cognitive distraction is not accompanied by body movements, which means that the visual modality data features are not evident. Therefore, relying solely on the visual modality is not sufficient to enhance the generalization performance of the model. In Table 5, the average recognition recall rate for slight risk status using the multimodal data fusion methods increases to 84.1%, and that of the proposed V-SFT is the highest at 90.2%, showing the performance enhancement achieved by modality fusion. Similarly, in some other works that rely on only the sensor modality, although typical risky vehicle states such as sharp acceleration or deceleration [20], lane departure [18], and close following distance [19] can be captured, methods based on only this single-modal data may often cause misjudgments. The resulting traffic accidents or low fuel efficiency will be detrimental to social health and

environmental sustainability. In the confusion matrix heatmaps in Figure 14, it can be observed that the methods that use only sensor-modal data largely confuse the two statuses of moderate risk and severe risk. The specificity of the driver's driving patterns [21] and the frequent vehicle starts and stops on certain road sections may have an impact on the recognition accuracy. However, it can be clearly seen in Figure 16 that the multimodal data fusion method can better solve this problem, and the proportion of misjudgments is greatly reduced, which once again reveals that the complementary modal information strengthens the sustainability understanding of driving status.

There are few existing studies on driving status recognition by fusing multimodal data. In fact, different fusion strategies lead to differences in the extensibility and operability of the models, which determines whether it can adapt to the sustainability of the future development of autonomous driving technology and intelligent transportation systems. In terms of sustainable expansion capability, previous studies used traditional convolutional backbone networks such as VGG [53] and AlexNet [54] or improved convolutional backbone networks such as multiscale attention CNN [55] to extract the spatial information of driver actions and used traditional recurrent backbone networks such as LSTM [51] or improved recurrent backbone networks such as attention-based GRU [50] to learn the temporal vehicle motion information. Although these methods take advantage of different types of networks to process corresponding data modalities, the mechanical combination of different network branches ignores the spatiotemporal dependence of multimodal features and limits the sustainable expansion capability of the model. The reason is that non-shareable network parameters need to be trained for different backbones. Additionally, with the addition of more modal data, such as acoustic audio [28], it may be necessary to adjust the entire model structure for compatibility. The transformer backbone network used in this study has proven to be applicable in many fields, such as computer vision, time series classification, and natural language processing [42]. This means that only one unified backbone network can adapt to multiple modal data, which can solve the problem of the insufficient extensibility of existing models. In terms of operationality, some studies required training the models in stages [27,30]. The approach was to obtain a set of driving status confidence probabilities based on the vision modality and another set of confidence probabilities based on the sensor modality. Then, the two sets were probabilistically fused through the Bayesian network [56]. The V-SFT in this study performed feature-level prefusion on different modalities, and the encoder block maximized the information interaction of the vision-sensor (human-vehicle) modality. In addition, this end-to-end structure is more integrated and concise in the process from data preparation and model training to deployment.

However, there are some limitations in this study. Due to the current limited experimental conditions, the recruited participants were generally school students, all of whom were young males with less driving experience. Future work will consider expanding the research sample to further explore the impact of variables such as gender, occupation, and driving experience. In addition, although the human–vehicle information fusion has been achieved, there are still failure or misjudgment scenarios for complex and changeable road traffic conditions. Therefore, an important direction for future work will be to integrate perceived or stored roadside scene information into the model on this basis. Additionally, data collection should also use real vehicles to conduct experiments on real roads as much as possible to simultaneously obtain human–vehicle–road collaborative information [57]. Last, despite the proposed V-SFT having excellent accuracy and operability, its real-time performance needs to improve. To use the V-SFT for real-time recognition, it is critical to further optimize the model parameters before the hardware deployment, and to ensure the data transmission rate and sufficient computing resources of the hardware.

5. Conclusions

Sustainable development requires balancing the interests of multiple aspects, such as the economy, environment, and society. In this study, an end-to-end vision-sensor fusion

transformer model, termed V-SFT, was constructed for recognizing distracted driving status risk levels. The model may help to improve transportation and energy efficiency, and its extensibility has great potential for future sustainable developments. It consisted of three main modules: early fusion of vision-modal data and sensor-modal data, modality information interaction in the encoder block, and classifier head for risk level inference. First, the tokenization process was built separately to extract tokens of different modalities, and then feature prefusion was employed to aggregate the respective modality representations. Second, after position encoding, the multimodal joint tokens continuously interacted based on a multi-head attention mechanism. Finally, the token at a specific position was extracted for risk level recognition in the classifier head. To verify the effectiveness of V-SFT, a data collection platform was developed to synchronously record driver behaviour visual signals and vehicle motion sensing signals during simulated driving. The V-SFT was evaluated under different modality inputs on the collected datasets. It was shown that V-SFT can outperform the other compared models regardless of whether only vision-modality or only sensor-modality input was used. When relying on the fusion input of two modalities, V-SFT achieved the best performance with an recognition accuracy of 92%. In future work, to obtain a more robust model, the scope of data collection will be further expanded. Benefiting from the sustainability of the model's expansion capability, more modal data will be collected from other types of sensors, and different regularization techniques will be applied to optimize the model and its inputs. The model could be integrated into advanced driver assistance systems (ADAS) as a separate module to provide a basis for the allocation or switching of control rights for intelligent vehicles.

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

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 51975172), and the Humanities and Social Sciences project of the Ministry of Education of China (Grant No. 19YJCZH005), and the Zhejiang Provincial Natural Science Foundation of China (Grant No. LY19E050012), and the National Natural Science Foundation of China (Grant No. 52175088).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by Ethics Committee of the Second Xiangya Hospital (No. 2018-S086).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets of AUC and State Farm used for pretraining in this study can be found at https://abouelnaga.io/projects/auc-distracted-driver-dataset/ and https://www.kaggle.com/competitions/state-farm-distracted-driver-detection. The dataset of UAH-DriveSet can be found at http://www.robesafe.uah.es/personal/eduardo.romera/uah-driveset/. The collected datasets were temporarily not publicly available due to privacy or ethical restrictions.

Acknowledgments: The authors are grateful to Xiexing Feng from University of Windsor for his assistance in conducting the analysis.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Hu, Y.; Qu, T.; Liu, J.; Shi, Z.Q.; Zhu, B.; Cao, D.P.; Chen, H. human-machine cooperative control of intelligent vehicle: Recent developments and future perspectives. *Acta Autom. Sin.* **2019**, *45*, 1261–1280.
2. Xian, H.; Hou, Y.; Wang, Y.; Dong, S.; Kou, J.; Li, Z. Influence of Risky Driving Behavior and Road Section Type on Urban Expressway Driving Safety. *Sustainability* **2022**, *15*, 398. [CrossRef]

3. Distracted Driving Statistics. 2022. Available online: https://www.bankrate.com/insurance/car/distracted-driving-statistics (accessed on 9 August 2022).
4. Rahman, M.M.; Islam, M.K.; Al-Shayeb, A.; Arifuzzaman, M. Towards sustainable road safety in Saudi Arabia: Exploring traffic accident causes associated with driving behavior using a Bayesian belief network. *Sustainability* **2022**, *14*, 6315. [CrossRef]
5. Sayed, I.; Abdelgawad, H.; Said, D. Studying driving behavior and risk perception: A road safety perspective in Egypt. *J. Eng. Appl. Sci.* **2022**, *69*, 22. [CrossRef]
6. Suzuki, K.; Tang, K.; Alhajyaseen, W.; Suzuki, K.; Nakamura, H. An international comparative study on driving attitudes and behaviors based on questionnaire surveys. *IATSS Res.* **2022**, *46*, 26–35. [CrossRef]
7. Wang, L.; Wang, Y.; Shi, L.; Xu, H. Analysis of risky driving behaviors among bus drivers in China: The role of enterprise management, external environment and attitudes towards traffic safety. *Accid. Anal. Prev.* **2022**, *168*, 106589. [CrossRef] [PubMed]
8. Ge, H.M.; Zheng, M.Q.; Lv, N.C.; Lu, Y.; Sun, H. Review on driving distraction. *J. Traffic Transp. Eng.* **2021**, *21*, 38–55.
9. Li, N.; Busso, C. Predicting perceived visual and cognitive distractions of drivers with multimodal features. *IEEE Trans. Intell. Transp. Syst.* **2014**, *16*, 51–65. [CrossRef]
10. Grahn, H.; Kujala, T. Impacts of touch screen size, user interface design, and subtask boundaries on in-car task's visual demand and driver distraction. *Int. J. Hum.-Comput. Stud.* **2020**, *142*, 102467. [CrossRef]
11. Horrey, W.J.; Lesch, M.F.; Garabet, A. Dissociation between driving performance and drivers' subjective estimates of performance and workload in dual-task conditions. *J. Saf. Res.* **2009**, *40*, 7–12. [CrossRef]
12. Sun, Q.; Wang, C.; Guo, Y.; Yuan, W.; Fu, R. Research on a cognitive distraction recognition model for intelligent driving systems based on real vehicle experiments. *Sensors* **2020**, *20*, 4426. [CrossRef] [PubMed]
13. Peng, Q.; Xu, W. Crop nutrition and computer vision technology. *Wirel. Pers. Commun.* **2021**, *117*, 887–899. [CrossRef]
14. Craye, C.; Karray, F. Driver distraction detection and recognition using RGB-D sensor. *arXiv* **2015**, arXiv:1502.00250.
15. Behera, A.; Keidel, A.; Debnath, B. Context-driven multi-stream LSTM (M-LSTM) for recognizing fine-grained activity of drivers. In Proceedings of the 40th Pattern Recognition German Conference (GCPR), Stuttgart, Germany, 9–12 October 2018; pp. 298–314.
16. Eraqi, H.M.; Abouelnaga, Y.; Saad, M.H.; Moustafa, M.N. Driver distraction identification with an ensemble of convolutional neural networks. *J. Adv. Transp.* **2019**, *2019*, 21–32. [CrossRef]
17. Xing, Y.; Lv, C.; Wang, H.; Cao, D.; Velenis, E.; Wang, F.Y. Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE Trans. Veh. Technol.* **2019**, *68*, 5379–5390. [CrossRef]
18. Son, J.; Park, M. Detection of cognitive and visual distraction using radial basis probabilistic neural networks. *Int. J. Automot. Technol.* **2018**, *19*, 935–940. [CrossRef]
19. Kountouriotis, G.K.; Wilkie, R.M.; Gardner, P.H.; Merat, N. Looking and thinking when driving: The impact of gaze and cognitive load on steering. *Transp. Res. Part F Traffic Psychol. Behav.* **2015**, *34*, 108–121. [CrossRef]
20. Osman, O.A.; Hajij, M.; Karbalaieali, S.; Ishak, S. A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. *Accid. Anal. Prev.* **2019**, *123*, 274–281. [CrossRef]
21. Lansdown, T.C. Individual differences during driver secondary task performance: Verbal protocol and visual allocation findings. *Accid. Anal. Prev.* **2002**, *34*, 655–662. [CrossRef]
22. Reimer, B. Impact of cognitive task complexity on drivers' visual tunneling. *Transp. Res. Rec.* **2009**, *2138*, 13–19. [CrossRef]
23. Ding, X.; Wang, Z.; Zhang, L.; Wang, C. Longitudinal vehicle speed estimation for four-wheel-independently-actuated electric vehicles based on multi-sensor fusion. *IEEE Trans. Veh. Technol.* **2020**, *69*, 12797–12806. [CrossRef]
24. Gao, L.; Xiong, L.; Xia, X.; Lu, Y.; Yu, Z.; Khajepour, A. Improved vehicle localization using on-board sensors and vehicle lateral velocity. *IEEE Sens. J.* **2022**, *22*, 6818–6831. [CrossRef]
25. Malawade, A.V.; Mortlock, T. HydraFusion: Context-aware selective sensor fusion for robust and efficient autonomous vehicle perception. In Proceedings of the 13th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS), Milano, Italy, 4–6 May 2022; pp. 68–79.
26. Alsuwian, T.; Saeed, R.B.; Amin, A.A. Autonomous Vehicle with Emergency Braking Algorithm Based on Multi-Sensor Fusion and Super Twisting Speed Controller. *Appl. Sci.* **2022**, *12*, 8458. [CrossRef]
27. Omerustaoglu, F.; Sakar, C.O.; Kar, G. Distracted driver detection by combining in-vehicle and image data using deep learning. *Appl. Soft Comput.* **2020**, *96*, 106657. [CrossRef]
28. Du, Y.; Raman, C.; Black, A.W.; Morency, L.P.; Eskenazi, M. Multimodal polynomial fusion for detecting driver distraction. *arXiv* **2018**, arXiv:1810.10565.
29. Craye, C.; Rashwan, A.; Kamel, M.S.; Karray, F. A multi-modal driver fatigue and distraction assessment system. *Int. J. Intell. Transp. Syst. Res.* **2016**, *14*, 173–194. [CrossRef]
30. Streiffer, C.; Raghavendra, R.; Benson, T.; Srivatsa, M. Darnet: A deep learning solution for distracted driving detection. In Proceedings of the 18th Acm/Ifip/Usenix Middleware Conference: Industrial Track, New York, NY, USA, 11–13 December 2017; pp. 22–28.
31. Abouelnaga, Y.; Eraqi, H.M.; Moustafa, M.N. Real-time distracted driver posture classification. *arXiv* **2017**, arXiv:1706.09498.
32. Alotaibi, M.; Alotaibi, B. Distracted driver classification using deep learning. *Signal Image Video Process.* **2020**, *14*, 617–624. [CrossRef]

33. Romera, E.; Bergasa, L.M.; Arroyo, R. Need data for driver behavior analysis? Presenting the public UAH-DriveSet. In Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 387–392.
34. Lv, N.C.; Zheng, M.F.; Hao, W.; Wu, C.Z.; Wu, H.R. Forward collision warning algorithm optimization and calibration based on objective risk perception characteristic. *J. Traffic Transp. Eng.* **2020**, *20*, 172–183.
35. Bowden, V.K.; Loft, S.; Wilson, M.D.; Howard, J.; Visser, T.A. The long road home from distraction: Investigating the time-course of distraction recovery in driving. *Accid. Anal. Prev.* **2019**, *124*, 23–32. [CrossRef]
36. Chen, H.; Cao, L.; Logan, D.B. Investigation into the effect of an intersection crash warning system on driving performance in a simulator. *Traffic Inj. Prev.* **2011**, *12*, 529–537. [CrossRef] [PubMed]
37. Chen, H.; Liu, H.; Feng, X.; Chen, H. Distracted driving recognition using Vision Transformer for human-machine co-driving. In Proceedings of the 5th CAA International Conference on Vehicular Control and Intelligence (CVCI), Tianjin, China, 29–31 October 2021.
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
39. Sirignano, J.; Spiliopoulos, K. Scaling limit of neural networks with the Xavier initialization and convergence to a global minimum. *arXiv* **2019**, arXiv:1907.04108.
40. Phaisangittisagul, E. An analysis of the regularization between L2 and dropout in single hidden layer neural network. In Proceedings of the 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), Bangkok, Thailand, 25–27 January 2016; pp. 174–179.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 72–82.
43. Yu, B.; Bao, S.; Zhang, Y.; Sullivan, J.; Flannagan, M. Measurement and prediction of driver trust in automated vehicle technologies: An application of hand position transition probability matrix. *Transp. Res. Part C Emerg. Technol.* **2021**, *124*, 102957. [CrossRef]
44. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Adam, H. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
45. Wang, C.; Chen, D.; Hao, L.; Liu, X.; Zeng, Y.; Chen, J.; Zhang, G. Pulmonary image classification based on inception-v3 transfer learning model. *IEEE Access* **2019**, *7*, 146533–146541. [CrossRef]
46. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. CMT: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185.
47. Banerjee, T.P.; Das, S. Multi-sensor data fusion using support vector machine for motor fault detection. *Inf. Sci.* **2012**, *217*, 96–107. [CrossRef]
48. El Haouij, N.; Poggi, J.M.; Ghozi, R.; Sevestre-Ghalila, S.; Jaïdane, M. Random forest-based approach for physiological functional variable selection for driver's stress level classification. *Stat. Methods Appl.* **2019**, *28*, 157–185. [CrossRef]
49. Alvarez-Coello, D.; Klotz, B.; Wilms, D.; Fejji, S.; Gómez, J.M.; Troncy, R. Modeling dangerous driving events based on in-vehicle data using Random Forest and Recurrent Neural Network. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 165–170.
50. Javed, A.R.; Ur Rehman, S.; Khan, M.U.; Alazab, M.; Reddy, T. CANintelliIDS: Detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 1456–1466. [CrossRef]
51. Khodairy, M.A.; Abosamra, G. Driving behavior classification based on oversampled signals of smartphone embedded sensors using an optimized stacked-LSTM neural networks. *IEEE Access* **2021**, *9*, 4957–4972. [CrossRef]
52. Yuan, Y.; Lin, L.; Liu, Q.; Hang, R.; Zhou, Z.G. SITS-Former: A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102651. [CrossRef]
53. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
54. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
55. Hu, Y.; Lu, M.; Lu, X. Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural network. *Signal Process. Image Commun.* **2020**, *81*, 115697. [CrossRef]
56. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **1997**, *29*, 131–163. [CrossRef]
57. Wu, J.; Kong, Q.; Yang, K.; Liu, Y.; Cao, D.; Li, Z. Research on the Steering Torque Control for Intelligent Vehicles Co-Driving with the Penalty Factor of Human–Machine Intervention. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *53*, 59–70. [CrossRef]