EmoTake: Exploring Drivers' Emotion for Takeover Behavior Prediction

Yu Gu*, *Senior Member, IEEE*, Yibing Weng*, Yantong Wang, Meng Wang, Guohang Zhuang, Jinyang Huang, Xiaolan Peng, Liang Luo, Fuji Ren, *Senior Member, IEEE*

Abstract—The blossoming semi-automated vehicles allow drivers to engage in various non-driving-related tasks, which may stimulate diverse emotions, thus affecting takeover safety. Though the effects of emotion on takeover behavior have recently been examined, how to effectively obtain and utilize drivers' emotions for predicting takeover behavior remains largely unexplored. We propose EmoTake, a deep learning-empowered system that explores drivers' emotional and physical states to predict takeover readiness, reaction time, and quality. The key enabler is a deep neural framework that extracts drivers' fine-grained body movements from a camera and interprets them into drivers' multi-channel emotional and physical information (e.g., facial expression, and head pose) for prediction. Our study (N = 26) verifies the efficiency of EmoTake and shows that: 1) facial expression benefits prediction; 2) emotions have diverse impacts on takeovers. Our findings provide insights into takeover prediction and in-vehicle emotion regulation.

Index Terms—Affect Analysis, Decision-making, Unobtrusive, Dataset, AI-human Interaction, Takeover.

1 Introduction

The blossoming semi-driving technology envisions a future where drivers are no longer bounded by the steering wheels and can engage in various non-driving-related tasks (NDRTs) such as reading or watching videos on mobiles. While these rich activities greatly enhance drivers' experience, they could also change drivers' physical state and stimulate drivers' diverse emotions, thus affecting takeover safety. Specifically, drivers are demanded to take over the vehicles from time to time, due to legal requirements or technology limitations. As such, takeover requests (TORs) are generated and broadcast in the form of auditory, visual, or vibrotactile warnings by the vehicles [1]–[3]. However, TORs are not always timely responses, especially when drivers are emotionally engaged in various NDRTs. Even if TORs are promptly answered, there is no guarantee that drivers will handle the upcoming situation properly (e.g., immersing in a comedy and overreacting to sudden TORs), raising takeover safety concerns.

*Equal contribution. Yu Gu proposed leveraging emotions for takeover behavior prediction and designed the framework. Yibing Weng implemented the framework and conducted the experiments. Yu Gu, Yibing Weng, Liang Luo (corresponding author), and Fuji Ren are with I^+ Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China; Yantong Wang (corresponding author), Meng Wang, and Guohang Zhuang are with the School of Computer and Information, Hefei University of Technology, China.
Email: yugu.bruce@ieee.org

To address such concerns, online takeover behavior prediction has recently been explored to help the vehicle foresee how a driver performs in takeovers [4], [5]. In general, they leveraged data mining on drivers' physical state data with various driving-related contexts (e.g., vehicle [4], traffic [6], and weather [7]) to predict different aspects of takeover behaviors such as takeover readiness [8] and reaction time [7]. However, such prediction models are still insufficient in application since takeover is a complex task involving attention, information perception, real-time judgment, and decision execution [9], all affected by emotions. But little explicit consideration of the driver's emotional state has been given in prediction, despite its effects on takeover behavior recently being examined [10]–[12].

To this end, we propose **EmoTake** (see Fig. 1), a deep learning empowered system that explores drivers' emotions to help forecast: (1) *takeover readiness* - whether the driver is ready for a TOR [8]; (2) *takeover reaction time* - how long it takes for the driver to resume manual driving after a TOR [13]; (3) *takeover quality* - how well the driver handles a takeover [14]. Instead of multiple costly and obtrusive apparatuses as in prior studies, **EmoTake** leverages a single camera for contactless human body-related data collection. The key enabler underlying **EmoTake** is a deep neural framework that extracts drivers' fine-grained body movements (i.e., face, blood vessel, eye, head, and upper body) from time-series images captured by the camera and interprets them into drivers' multi-channel emotional and physical information (i.e., facial expression, eye movement, head pose, and body posture) for the prediction with vehicle data. We realize **EmoTake** in a driving simulator and validate its performance with a human-subject study involving 26 participants. After pre-driving questionnaires and simulator familiarization, they were put under L3 automation [14] while watching video clips as the NDRT before encountering TORs. These clips could induce emotions that are commonly experienced like boredom, anger, and excitement. Empirical results have verified the efficiency of **EmoTake** in predicting takeover readiness, reaction time, and quality with an accuracy of 91.74%, 88.55%, and 81.71%, respectively.

With this work, we aim to encourage designers to consider vision as an effective signal for predicting drivers' takeover behavior. The main contributions of our work are:

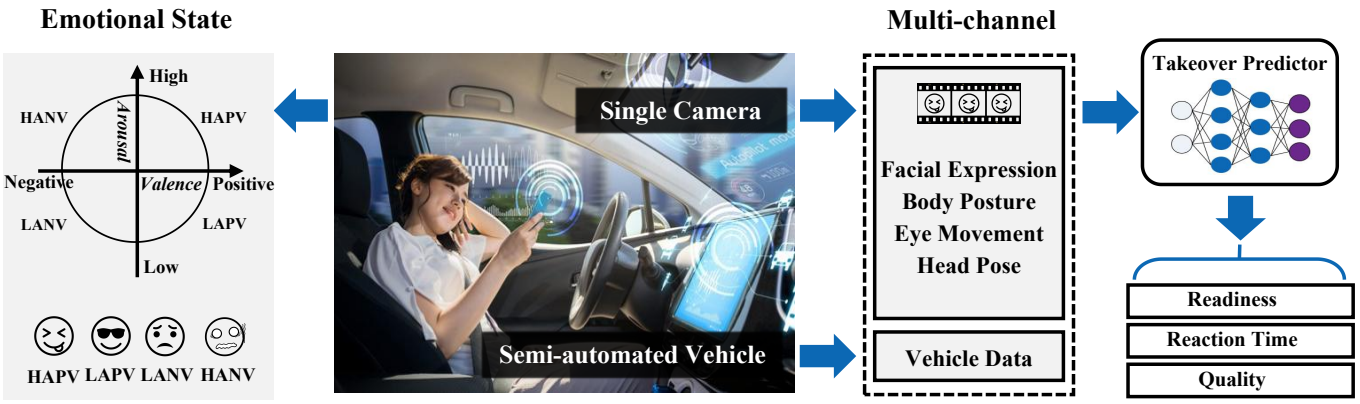


Fig. 1. EmoTake explores drivers' emotional and physical states for takeover prediction. A deep neural framework is designed to extract drivers' fine-grained body movements from a single camera and interpret them into drivers' multi-channel emotional and physical information (i.e., facial expression, body posture, eye movement, and head pose) for predicting takeover behavior in readiness, reaction time, and quality with vehicle data.

- The first system that explicitly encodes drivers' facial expressions for predicting takeover behaviors.
- Empirical evidence of facial expression benefiting takeover behavior prediction.
- Correlation analysis of driver's emotion on takeover behaviors. Some findings are contrary to manual driving, e.g., positive valence implies higher collision risks while high arousal leads to gentler steering.
- Suggestions for designers who are interested in takeover behavior prediction and in-vehicle emotion regulation.

2 Related Work

2.1 Emotion in Driving

Manual driving is a complex task that could affect drivers' emotions when they are interacting with passengers, the external environment, and other road users [9]. Emotions in manual driving have long been explored in many aspects [15]–[17]. In these studies, two emotion representations have frequently been used [18]. One is the discrete model, where emotions are discrete and fundamentally different constructs like anger, fear, and happiness. The other is the dimensional model where emotions can be characterized on a dimensional basis in groupings, e.g., the well-recognized valence-arousal (VA) model. Here valence refers to how negative or positive a stimulus is, while arousal indicates how sleeping or exciting a stimulus is, respectively [19].

For the discrete model, anger receives wide attention since it leads to risky behaviors such as speeding and traffic rule violations [20], [21]. Also, it lowers the perceived safety of drivers and thus degrades their driving performance [22]. As a result, it becomes one of the most significant contributors to fatal crashes [23] (e.g., increasing the risk of a crash by 9.8 times [24]). Besides anger, other commonly experienced discrete emotions in driving like happiness, sadness, fear, and boredom (fatigue) have also been studied [22], [25]–[27]. Their impacts on various aspects like risk perception, response, and steering have been examined.

The dimensional model allows researchers to conduct not only qualitative analysis but also quantitative analysis on how emotion affects driving. Prior studies show that negative valence, stimulated via words [28] or images [29], distracts drivers and leads to higher speeds and worse lateral control [30]. On the other hand, high arousal, provoked via images [29] or musical tempos [31], sharpens drivers' attention to achieve a faster hazard response [32]. However, it is also shown to be related to radical driving [32]. In general, previous research suggests associations between positive valence and better vehicle control, higher arousal, and faster risk response, respectively.

Despite how emotion affects manual driving has been well studied, how emotion affects takeover remains unexplored until very recently. N. Du *et al.* [10] presented the first empirical study in a driving simulator where participants experienced takeovers under L3 automation while watching movie clips for emotion induction. Their analysis suggests that drivers in positive valence tend to make a smaller acceleration and jerk when re-taking control, leading to better driving quality. They also show that high arousal does not essentially lead to shorter takeover time, contrary to the observation in manual driving. Their work provides critical insights into the role of emotions in takeovers and inspires us to explicitly consider emotion as a major factor in predicting driver takeover behavior.

2.2 Takeover Behavior Prediction

Takeover safety draws increasing attention with the development of autonomous driving technologies. The key challenge is how to ensure a smooth transition of controls from vehicles to drivers, who could be immersed in NDRTs and become incapable of driving. One possible way is to study how different factors such as the external driving environment [35], [36], types of NDRTs [37]–[39], individual characteristics [38], [40], [41], and human-machine interface [42], [43] affect takeover behaviors, and then set up offline regulations for precaution. However, precautions will not help vehicles to foresee whether a particular driver can handle a particular takeover event well.

Therefore, online prediction of drivers' takeover behavior via computational models has attracted much attention quite recently. In general, they leveraged data mining on various driving contexts (e.g., driver [44], vehicle [4], and

TABLE 1
Literature review on takeover performance prediction.

Ref.	Takeover Metric	User-context		Vehicle-context	Unobtrusive	Setting	Modalities	Method	Participant
		Emotional	Physical						
[33]	Takeover time	×	√	√	√	Simulator	Eye movement Body posture	Machine learning	88 subjects
[8]	Readiness	×	√	√	√	Simulator	Head movement Eye movement	Machine learning	81 subjects (45M & 36F)
[6]	Readiness	×	√	√	√	Real	Gaze Pose Hand activity Foot activity	CNN LSTM	11 subjects (7M & 4F)
[7]	Reaction time	×	√	√	×	Simulator	GSR Head pose Gaze data Eye blink Heart rate	Machine learning	102 subjects (62M & 40F)
[5]	Takeover time	×	√	√	Not mentioned	Simulator & Real from dataset [34]	Hand posture Cognitive load	XGBoost	4556 subjects [34] (from 129 studies)
[4]	Intention Reaction time Quality	×	√	√	×	Simulator	GSR Heart rate Eye movement	Deep learning network	20 subjects (9M & 11F)
Ours.	Readiness Reaction time Quality	√	√	√	√	Simulator	Encoded FE Head pose Body posture Eye movement	iTransformer	26 subjects (14M & 12F)

GSR: Galvanic skin response; CNN: convolutional neural network; LSTM: Long short-term memory; XGBoost: eXtreme Gradient Boosting; FE: Facial Expression

traffic [33]) to predict different aspects of takeover behaviors. These models enable vehicles to continuously monitor the drivers and make real-time judgments to ensure a safe takeover. Table 1 provides a detailed review of some state-of-the-art approaches, and compares them with our work for a clear positioning. As suggested in [45], we deliberately omitted the prediction accuracy achieved by different models, to avoid any superiority misinterpretation. Because their experimental conditions are fundamentally different, and direct comparison of the prediction accuracy is meaningless.

Lotz and Weisenberger [33] chose drivers' eye movement and body posture data collected via two devices (Smart Eye Pro and Microsoft Kinect) as input in a simulator study to predict drivers' takeover time (classified into four classes). They used a linear support vector machine (SVM) for data mining. Braunage *et al.* [8] collected drivers' eye movement and head movement data via three cameras in a high-end Mercedes-Benz driving simulator, and explored them with NDRT involvement data in a SVM classifier to forecast takeover readiness (high and low). Deo and Trivedi [6] leveraged 5 cameras to capture drivers' gaze, pose, and hand and foot activities in an extensive naturalistic study of a conditionally autonomous vehicle running on Californian freeways. They used a Long Short Term Memory (LSTM) model to estimate the drivers' takeover readiness index on a 5-point scale. Ayoub *et al.* [5] explored the eXtreme Gradient Boosting (XGBoost) algorithm on a hybrid dataset gathered from 129 studies in both real-world scenarios and driving simulators to predict takeover time. Besides vehicle data, they used drivers' hand posture and cognitive load.

Recently, there has been a trend of exploring drivers' physiological data to enrich the understanding of drivers' physical states. N. Du *et al.* [7], [44] monitored drivers' eye movement, heart rate, and galvanic skin response (GSR) via Tobii Pro-Glasses 3 and Shimmer3 GSR+, respectively. They explored these drivers' data with a random forest classifier to predict takeover quality in two classes (good and bad). With the same devices, Pakdamanian *et al.* [4] present a deep neural network to handle drivers' data with the vehicle and NDRT data for predicting takeover intention, time, and quality. In these work, more data modalities usually lead to more apparatuses, which are costly and sometimes obtrusive.

In summary, despite the fact that emotion plays an important role in driving and its impacts on takeover behaviors have recently been examined, none of the previous works has considered drivers' emotions for the prediction of takeover behavior. Our preliminary study has revealed the potential of leveraging emotions to enhance takeover safety via a distributed Deep Neural Network (DNN) mounted separately on the cloud and vehicle end [46]. EmoTake further pushes along this direction and explicitly explores drivers' encoded facial expressions to help predict driver takeover behavior via an inverted Transformer-based deep forecasting network handling the multi-channel time series data. Moreover, EmoTake has been systematically evaluated to provide a better understanding of how different data channels matter in prediction and how different emotions affect takeover behavior. Our correlation analysis of driver's emotions on takeover behaviors provides interesting findings, some of which are contrary to manual driving, and provide aids for researchers and engineers who are interested in takeover behavior prediction and in-vehicle emotion regulation.

3 METHODOLOGY

EmoTake roots in a deep neural framework that extracts drivers' fine-grained body movements and interprets them into drivers' multi-channel emotional and physical information for prediction. Fig. 2 shows the workflow of the EmoTake, which consists of three parts: data collection, data labeling, and takeover behavior prediction. The specific details of these three parts are introduced as follows.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3399328

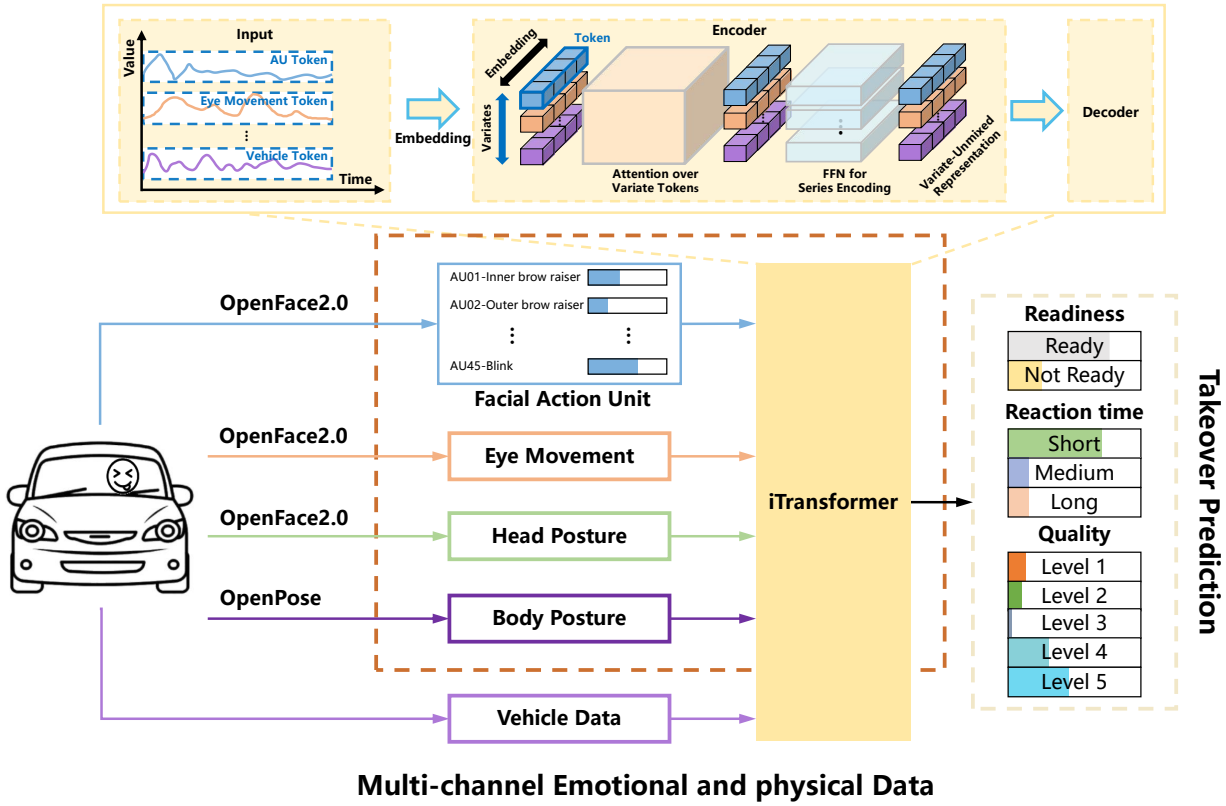


Fig. 2. The proposed framework for extracting drivers' multi-channel emotional and physical information as well as predicting takeover behaviors. Specifically, we utilize a series of cutting-edge deep learning methods to interpret drivers' fine-grained body movements into multi-channel information, i.e., facial expression in Action Units (AUs), eye movement, head pose, and body posture. Later, we combine these channels with vehicle data via multiple neural networks to predict takeover behaviors from three aspects: readiness, reaction time, and quality.

3.1 Data Collection

EmoTake leverages a camera to collect drivers' facial expressions, body posture, eye movement, and head pose data in an unobtrusive way. Below we introduce details on each channel.

3.1.1 Encoded Facial expression

Facial expression constitutes a primary means of emotional expression [47], [48]. Drivers engaged in Non-Driving Related Tasks (NDRTs) may experience a range of emotions when facing TORs. To effectively capture these emotional nuances, we employ the Multi-task Convolutional Neural Network (MTCNN, [49]) to extract facial expression features from a camera, whose sampling frequency is set to 30 frames per second. Then, we downsample the footage to 15 frames per second (with one frame interval). As a result, the input of MTCNN is $(15 \times 10, 112 \times 112 \times 3)$.

We choose the prevailing Action Units [50], which represent the basic muscular movement that can be combined to form complex expressions, to encode the cropped facial images via OpenFace 2.0. Each AU has a specific code that represents a specific facial muscle group. OpenFace extracts 18 AUs in AU01_c and AU01_r. The former is a binary value indicating whether this AU is present, while the latter ranges from 0 (none) to 5 (maximum) denoting its intensity.

3.1.2 Body posture

Body posture can serve as a direct indicator of driving attention. By capturing the posture changes of the driver's upper body, various information attention-related tasks can be realized, e.g., fatigue detection [51]. EmoTake uses the pre-trained OpenPose model (see [52]) to detect limbs in frame images, providing 12-point coordinates in the upper limb region for each frame.

3.1.3 Eye movement and Head pose

Studies have demonstrated that eye movement and head pose are strongly associated with driving attention [53], [54]. In particular, the driver's focus, which can be assessed through their eye movement and head pose, clearly differs between distracted and non-distracted situations. The OpenFace 2.0 [55] toolbox is proven to be effective in extracting eye movement and head pose from video images. Thus, EmoTake employs this toolbox to pre-process the video data to obtain the eye movement features and the head pose features from each frame. The specific features of the eye movement extracted by OpenFace 2.0 are the direction of the left eye gaze (vector), the right eye gaze (vector), the angle of both eyes and the coordinates of the key points of the eye area in 2D and 3D. The specific features of the eye movement extracted by OpenFace 2.0 include the direction of the left eye gaze (vector), the direction of the right eye gaze (vector), the angle of both eyes, as well as the coordinates of key points in the eye area in both 2D and 3D. The total dimension of these eye movement features is $(10 \times 30, 288)$. In addition, the extracted features of the head pose include the position of the head relative to the camera

TABLE 2
List of extracted data used in EmoTake.

Data Source		Method	Data	Type	Size
Camera	Body Posture	OpenPose	Body key point coordinates	Int	(300, 12×2)
	Action Units	OpenFace 2.0	Facial action units	Float, Int	(300, 35)
	Eye movement		Eye gaze direction vector	Float	(300, 8)
			Eye region landmarks	Float	(300, 5×56)
	Head Posture		Head position	Float	(300, 6)
Vehicle	\	\	Speed	Float	(300, 1)
			Throttle pedal angle	Float	(300, 1)
			Brake pedal angle	Float	(300, 1)
			Steer wheel angle	Float	(300, 1)

TABLE 3
Descriptions of subjective quality ratings of takeover quality.

Takeover quality rating	Explanation
Level_1	Collisions of loss of control
Level_2	Endangerments of oneself or other road users (e.g., near misses)
Level_3	Occurrence of driving errors (e.g., late or insufficient braking)
Level_4	Imprecisions of vehicle control (i.e., imprecise lane keeping)
Level_5	Perfect performance (i.e., absence of imprecisions and errors)

and the rotation angles X, Y, and Z of the head. The total dimension of these head pose features is $((10 \times 30, 6)$.

3.1.4 Vehicle data

Vehicle data like speed, throttle pedal angle, brake pedal angle, and steering wheel angle are collected at a sampling rate of 30 Hz (See Table 2).

3.2 Data Labeling

Our supervised learning network needs labeled training data to make predictions. Below we introduce how we label the collected data.

- **Takeover readiness.** Takeover readiness is more of a subjective judgment and thus hard to decode from the perspective of a bystander [10]. Therefore, readiness is self-labeled into binary outcomes ("Ready" and "Not Ready") in our study as in [8].
- **Takeover reaction time.** It is defined as the duration between a TOR being issued and the takeover button being pressed on the steering wheel for control transition (see Fig. 5). To simplify the computation, we divide it into three categories. using the threshold derived from the distribution of empirical reaction time (see Fig. 6). In this way, the label is set to "Short" when reaction time $\in$(0, 1.22s], "Medium" when reaction time $\in$(1.22s, 2.10s], and "Long" when reaction time $\in$(2.10s, 3.5s] (See Section 4.6).
- **Takeover quality.** We choose the subjective measure of takeover quality to minimize the individual difference in the driving experience and habit as in the ISO/TR 21959-1 technical report [14]. Specifically, we select a 5-point scale for representing the different levels of quality rated by a trained external observer ("Level_1" = worst performance, "Level_5" = perfect performance). The explanations are listed in Table 3.

3.3 iTransformer Architecture for Takeover Behavior Prediction

As shown in Fig. 2, the physical and emotional input channels are time-series numerical data. Therefore, we present an iTransformer-based network, a recently proposed inverted transformer specifically designed for time series forecasting via effectively exploring the intra-modality and inter-modality correlations [56], to process the input data for takeover behavior prediction. Specifically, the current transformer-based networks are based on temporal tokens that are formed by fusing multiple channel data of the same timestamp. However, this unified timestamp embedding neglects the diverse syntax meanings of each channel, which could result in unaligned channel representations and meaningless attention maps. Taking our problem as an example, AUs and other physical channels like body posture have quite different meanings. To this end, iTransformer forms tokens on each channel in a certain look-back window instead of fusing different channels into one token at each time stamp, and then utilizes them via the attention mechanism to capture multi-channel correlations.

More specifically, We use a cross-entropy loss function and the Adam optimizer with a learning rate of 0.0001 to update network parameters in our 8-head 2-layer Transformer model. The model is trained for 100 epochs. The batch size is set to 32. The length of a token's look-back window on each channel is set to 10 seconds, corresponding to 300 frames ($10(s) \times 30(f/s)$). Tokens are then embedded into 512-dimensional vectors via a shared Linear Layer (300×512). The specific channel data is shown in Table 2. In the decoding layer, the output of the encoder passes through a GELU activation function, a dropout layer (dropout=0.1), and a linear layer (512×10), and finally outputs the classification result. We leverage the subject-independent cross-validation to eliminate potential interference like age and identity. Details about the losses during training and testing can be found in the *Appendix*.

TABLE 4
Details on participants.

Group	No. of Subjects	Mean age (years)	Std. age (years)	Mean Driving Exp.	No. of glass wearing
Male	14	26.4	5.4	3.0	13
Female	12	27.5	7	3.0	7
Age 20-30	20	24.1	2.6	1.9	16
Age 30-40	4	33.5	1.7	4.0	3
Age >40	2	42.0	1.4	11.0	1
All	26	26.9	6.1	2.9	20

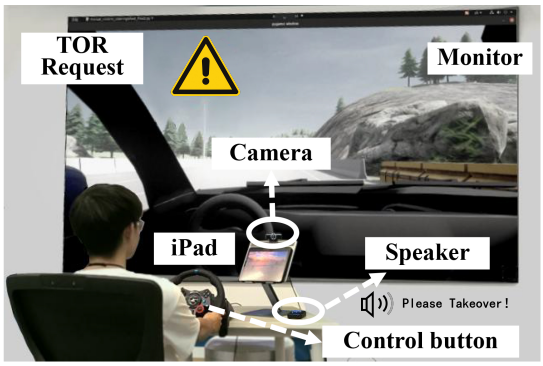


Fig. 3. Participant sitting in the driving simulator. The TOR signal consists of both sound ("Please takeover!") and image (a blinking exclamation mark)

4 User Study[1]

4.1 Participants

This study recruited 26 participants (14 males and 12 females) from our university, all of whom possess at least one year of driving experience. Their ages vary between 20 and 43 years, with an average of 26.9 years and a standard deviation of 6.1 years. All participants hold valid Chinese driving licenses and have an average driving experience of 2.9 years. Specific details can be found in Table 4.

Glasses could be challenging for OpenFace to decode facial expressions, e.g., blocking eyebrows. We have realized this problem when we are recruiting participants since the glass-wearing rate is particularly high in Chinese universities. We have taken several countermeasures to deal with this issue. Firstly, we carefully adjust the illumination in experiments to avoid light reflection on glasses, which is pretty problematic for OpenFace. Secondly, we manually check both the video footage and the recognition results of OpenFace after each experiment to ensure most of the AUs can be identified. Lastly, we conduct a careful search on the collected dataset after all the experiments have been done to remove noisy samples. As a result, we found out that OpenFace is a reliable AU detector on our dataset and works well in most cases.

4.2 Apparatus

We employed the *CARLA* software [57] to create the driving environment, which was presented on a 120-inch 4K resolution monitor (Changhong 90C7UG), as depicted in Fig. 3. Participants occupied a driving simulator seat equipped with a *Logitech G29* driving force racing wheel and floor pedals. Auditory and visual reminders for Takeover Requests (TORs) were provided through an external audio speaker and a pop-up window [1], [2]. An *Apple iPad Pro* was positioned on the right side of the driver at a 30-degree angle on an iron stand 20 cm above the dashboard for watching videos. We used a *Rapoo C260AF* webcam (30Hz, 720p, $28.7) to record drivers' facial expressions and behaviors during the experiments. The vehicle data, sampled at a rate of 30 Hz, was provided by *CARLA*.

4.3 Design

4.3.1 Emotion mapping into Valence-Arousal circumplex plane.

We selected the widely recognized valence-arousal model, illustrated in Fig. 4 (a), to characterize various emotions. In this model, valence indicates the degree of negativity or positivity of a stimulus, while arousal represents the level of intensity or excitement [45]. Compared to the traditional discrete model, this dimensional model allows us to conduct not only qualitative analysis but also quantitative analysis on how emotion affects driving, as shown later in Section 6.3.

4.3.2 Identifying emotion-inducing video clips.

We used video clips to induce emotions for two reasons:

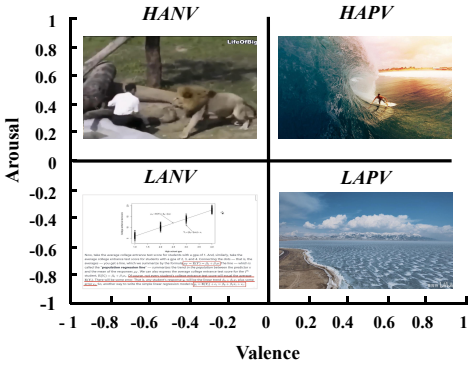
- It is one of the most common NDRTs appearing in the relevant literature [58].
- Its ability to stimulate diverse emotions has been effectively validated in previous studies [59]–[61].

Fig. 4 (b) summarizes the sources as well as the number of video clips we have selected for each quadrant on the valence-arousal plane. Our sources encompass a diverse range of content, including movies (such as Schindler's List), cartoons (like Tom and Jerry), news (like Elder abuse), shows (like Running Man), documentaries (like BBC Earthquakes), and online courses (like Mathematical Analysis). Each carefully selected clip is thoroughly vetted and approved by at least three trained reviewers. We give examples of each emotional quadrant in the *Appendix*, as well as the emotional status of the participants. More information about the videos can be found on our GitHub: https://github.com/yibingweng/EmoTake.

4.3.3 Takeover Requests.

Like in previous studies [4], we also chose obstacle avoidance as the criterion for takeover (see Fig. 5 for further

1. The Institutional Review Board (IRB) approval number is 106142023030728744 and the approving institution is University of Electronic Science and Technology of China.



. (a)

Emotion category	Number of video clips	Duration (min)	Source
HAPV (e.g., happy, exciting)	32	85.1	Tom and Jerry, Running man and etc.
HANV (e.g., fear, anger)	17	88.77	BBC earthquakes, Elder abuse news, Schindler's list, and etc.
LANV (e.g., bored, sad)	9	56.72	A series of mathematical analysis online course.
LAPV (e.g., relief, serene)	12	63.72	Beautiful trees and flowers in the world and etc.
Total	**70**	**294.31**	

. (b)

Fig. 4. (a) **Arousal** (1 = "High", -1 = "Low"), and **Valence** (1 = "Positive", -1 = "Negative"). ***HAPV***: high arousal and positive valence,; ***HANV***: high arousal and negative valence, ; ***LANV***: low arousal and negative valence; ***LAPV***): low arousal and positive valence. (b) The number, duration, and source of video clips for eliciting emotions.

information). TOR was delivered via an external audio speaker and a pop-up window as the auditory and visual alerts (see Fig. 3).

4.4 Measure

After the emotional induction, participants were required to complete the Self-Assessment Manikin (SAM) to assess their emotional state, while the takeover behaviors were evaluated by one experimental observer (see Section 4.6).

4.5 Procedure

We designed the experimental procedure (Our experiments were conducted by a protocol approved by the Institutional Review Board (IRB)) as shown in Fig. 5, which consisted of the following five steps:

4.5.1 *Briefing.*

Upon arrival at the experimental site, participants were asked to sign a consent form and complete a driving history questionnaire.

Before signing the consent form, we will introduce participants to the specifics of the experiment: All participants will simulate semi-autonomous drivers, and in good road conditions, the vehicle will drive itself along the road, while the driver will watch a video on an iPad. When the vehicle detects an obstacle, the driver will be alerted to take over by sounding the speaker and a warning on the monitor. The driver takes back control by pressing a control button to handle the task.

We asked participants to avoid high-risk driving behaviors as much as possible to approximate driving in a real

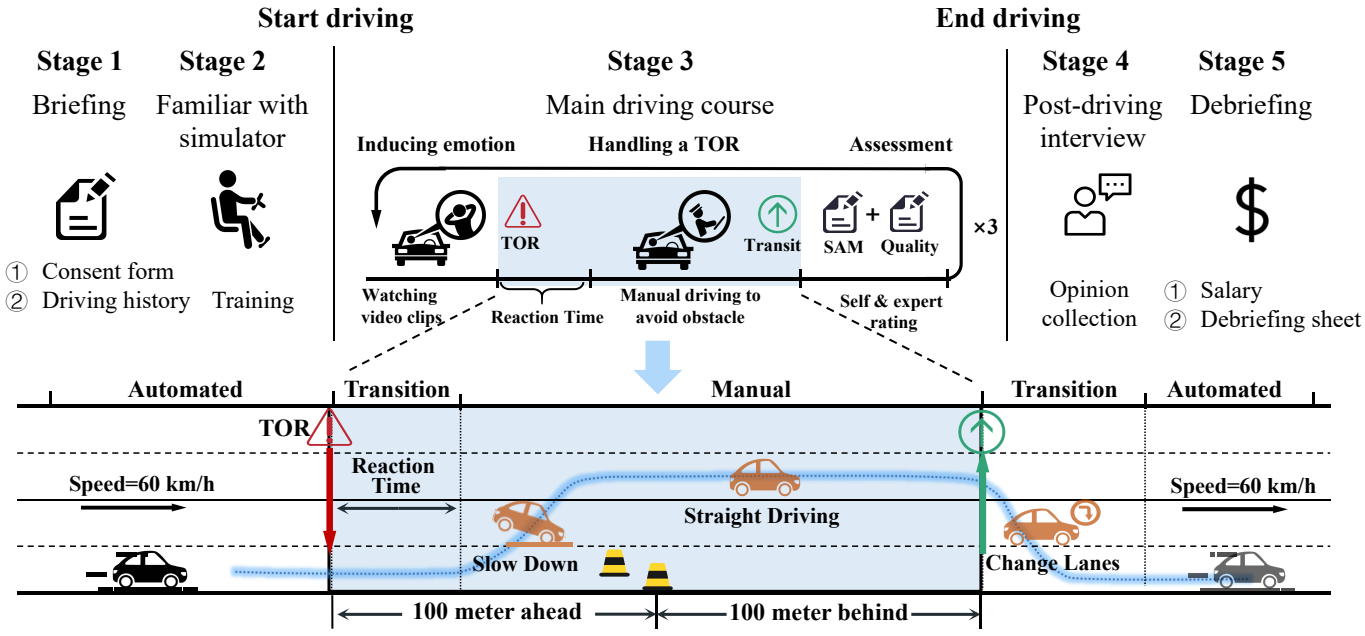


Fig. 5. The timeline of the user study procedure. The top graph shows a schematic view of an example of a completed procedure in our study, including stages 1-5: briefing, familiarization with the driving simulator, main driving course, post-driving interview, and debriefing. The bottom graph indicates the takeover timeline, including automated driving, encountering a TOR, resuming manual driving to avoid obstacles, and back to automated driving. The black car indicates automated driving, while the orange car means manual driving.

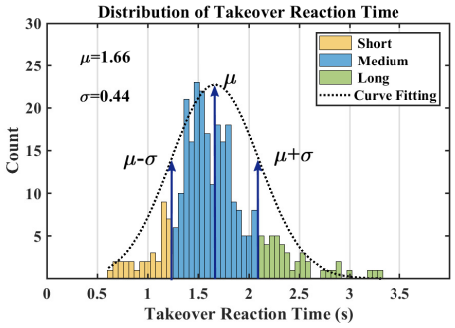


Fig. 6. Distribution of takeover reaction time using a bin width of 0.05s. Each data is fitted into the Normal Distribution (μ = 1.66, σ = 0.44). We use the μ-σ, μ+σ as threshold values to split the reaction time into "Short" (yellow), "Medium" (blue), and "Long" (green), respectively.

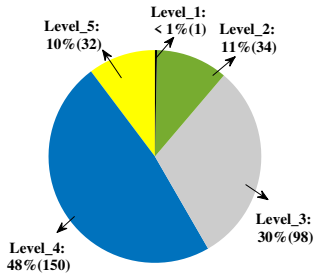


Fig. 7. Percentage of takeover quality level. We classify the takeover quality into 5 levels, ranging from "Level_1" to "Level_5".

environment. In the meantime, we encouraged them to be immersed in the videos as much as possible, to fully arouse their emotional feelings. Participants will be informed that the entirety of the experiment will be videotaped, as well as the 10 seconds of data before the takeover will be collected as our database. Regardless of whether a collision occurs or not, we will be paid normally upon completion of the experiment.

4.5.2 *Familiarization with the driving simulator.*

Every participant became familiar with the driving simulator by completing a 6-minute driving training, where they practiced lane changes, manual to automated driving activation/resumption, and achieved a common level of familiarity with the setup, NDRT, and auditory signals pitch. Prior to driving, participants were informed of the emotional dimensions (VA) to be explored during the current experiment and were therefore allowed to swipe back and forth across the screen of the *Apple iPad Pro* to identify appropriate video clips for inducing emotion. Once the participants confirmed that they were familiar enough with the simulators, they were told to rest for 5 minutes to ensure they could start the experiment in a relaxed manner.

4.5.3 *Main driving course*

After the familiarization session, participants were required to complete four driving courses, each containing three takeover events. Each course corresponds to one emotional quadrant, where participants were instructed to select and view video clips designed to simulate emotions in that quadrant. Furthermore, one takeover event can be roughly divided into three steps, namely inducing emotion by watching video clips, handling TORs to avoid obstacles, emotional state self-assessment, and takeover quality annotation. The specifics are outlined as follows:

4.5.4 *Post-driving interview.*

The experiment concluded with a brief semi-structured interview to gather further insights into the participant's emotions. The interview was recorded and centered on the following questions:

- How did the video clips make you feel?
- Which emotional quadrants affect your takeover behaviors and how?

As shown in Fig. 5, each course began with a command to activate the automated driving mode.

- Inducing emotion through video clips: While the vehicle was in automated driving mode, traveling at a speed of 60 km/h, participants watched emotional inducement video clips. We noticed that it typically took 1-3 minutes for participants to fully immerse themselves in the designated emotional quadrant. Therefore, we set TORs (Takeover Requests) to appear after 4 minutes of video watching. Participants were instructed to focus on the video clips displayed on the Apple iPad Pro until a TOR appeared.
- Handling TORs to avoid obstacles: As depicted in the bottom graph of Fig. 5, the TOR was positioned 100 meters ahead of the obstacle. When the TOR was issued, an audio speaker was activated and a warning interface appeared, prompting participants to take control. Participants then pressed a button on the steering wheel, slowed down the vehicle to avoid the obstacle, and continued driving to keep the vehicle in the current lane until 100 meters behind the obstacle. They then drove the vehicle into the fourth lane (ensuring consistent lane maintenance before the next TOR for further analysis) and pressed the same button to reactivate automated driving.
- Self-assessment of emotional state and annotation of takeover quality: Immediately following each takeover, participants were asked to recall the scenes from the video clips and assess their emotional state by completing the SAM questionnaire. Additionally, an experimental observer annotated the quality of each takeover.

The order of emotions being induced was counterbalanced via a 4×4 balanced Latin Square across participants to minimize the ordering effect.

4.5.5 *Debriefing.*

After the end of the interview, participants received a Debriefing Sheet and 50 dollars for their participation. Overall, the study lasted about 70 minutes.

4.6 Labels

We analyzed the collected data on takeover readiness, reaction time, and quality to get them properly labeled. In the

study, each participant finished 3 takeover events under one emotional quadrant, therefore, there existed 3 (takeovers) × 26 (participants) = 78 takeover records containing emotional state self-assessments and takeover quality annotations. In total, 4 (quadrants) × 78 (takeovers per quadrant) = 312 takeover records were collected.

Takeover readiness was defined as whether or not the participant was prepared enough to take over the vehicle. We labeled the takeover readiness as "Not Ready" and "Ready", respectively. In our study, "Not Ready" takes about 74.9%, while the rest was labeled as "Ready".

We plot the takeover reaction time with a grouped step size of 0.05s in Fig. 6 and find that its distribution fits a Normal Distribution (μ = 1.66, σ = 0.44), which was indicated in the black dashed curve line. Therefore, we divide the reaction time data into three categories, i.e., "Short", "Medium", and "Long", respectively. "Short" was defined as reaction time $\in$(0, μ-σ], "Medium" was defined as reaction time $\in$(μ-σ, μ+σ], and "Long" was defined as reaction time $\in$(μ+σ, 3.5s]. Fig. 6 leverages yellow, blue, and green to represent "Short", "Medium" and "Long", respectively.

As we described earlier in Section 3.2, we classified takeover quality into 5 levels. As illustrated in Fig. 7, 48% of takeover records were rated in "Level_4", which means smooth takeover behaviors. By contrast, 42% of takeover records were rated below "Level_3", which indicates imprecise or even dangerous takeover behaviors.

5 Performance Evaluation

Here we evaluate the performance of EmoTake on the data collected in the user study. The subject-independent cross-validation (5-fold) is used to eliminate potential interference like age and identity.

5.1 Metrics

Four metrics, i.e., $accuracy$, $precision$, $recall$, and $weighted\ F1-score$, are employed by EmoTake to evaluate the system performance, which can be represented as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Weighted\ F1\ Score = \sum 2 \times \frac{Precision \times Recall}{Precision + Recall} \times W_i \tag{4}$$

where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively. $Precision$ denotes the proportion of true positive examples in the prediction result. The $Recall$ represents the proportion of all positive examples that are correctly predicted. W_i represents the weight of the predicted category, where i means the number of categories.

5.2 How EmoTake Performs?

The accuracy, precision, recall, and Weighted F1-score results of iTransformer are shown in Table 5. Since the weighted F1-score is more capable of evaluating the imbalanced classification, we mainly use accuracy and weighted F1-score as evaluation metrics in the analysis.

TABLE 5
Comparison of EmoTake performance in different evaluation indicators.

Prediction	Accuracy	Precision	Recall	Weighted F1-score
Readiness	91.74	91.87	92.56	90.43
Reaction time	88.55	89.1	88.27	87.26
Quality	81.71	82.16	83.67	79.49

Table 6 shows the prediction results of EmoTake in the human-subject study. We notice that EmoTake is pretty accurate for takeover readiness, i.e., achieving a 91.74% accuracy and a 0.90 weighted F1-score. EmoTake also works fairly well in predicting takeover time and quality, i.e., obtaining a 88.55% accuracy and a 0.87 weighted F1 score for takeover time, while getting a 81.71% accuracy and a 0.79 weighted F1-score for takeover quality, respectively. The results have confirmed EmoTake's capability in predicting how a specific driver performs in a particular takeover event..

To further investigate the impact of emotion in different quadrants on the prediction performance, we separate the experimental results in each quadrant and conclude them in Table 6. Please note that: 1) our statistics are based on self-report emotions; 2) The number of samples in each emotional quadrant is different, so the average accuracy is based on their weighted sums.

We notice that EmoTake achieves the most satisfactory results for the ***HAPV*** quadrant (e.g., happy and exciting), with a 92.89% accuracy in takeover readiness prediction, a 89.83% accuracy in reaction time, and an 83.33% accuracy in takeover quality. But for takeover reaction time, the ***HANV*** has the best F1-score, i.e., 0.89. A possible and intuitive explanation is that high arousal leads to intensive body movements, affecting both emotional and physical channels and thus easier to recognize by the network. Another evidence to support this claim is that the lowest prediction performance is always located in the two low arousal states, i.e., ***LANV*** and ***LAPV***. For example, ***LANV*** (e.g., bored and sad) has the lowest readiness prediction as well as the quality prediction on accuracy, while ***LAPV*** (e.g., relief and serene) is the lowest on takeover reaction time.

5.3 How Different Channels Matter in Prediction?

To address this question, we conduct an ablation experiment on each channel, and Table 7 shows how EmoTake works without one specific channel. Firstly, we notice the all-fusing method always yields the best performance in the prediction of readiness, reaction time, and quality, indicating that every channel matters in prediction. Secondly, we notice that EmoTake suffers worse performance degeneration if removing the facial expression and eye movement channels than others, but the difference is not significant.

5.4 Naive Comparison

A naive comparison was conducted between the present iTransformer architecture and the Deep Neural Network

TABLE 6
Performance evaluation of EmoTake in different Valence-Arousal emotional quadrants.

V-A emotional quadrants	Takeover readiness		Takeover reaction time		Takeover quality	
	Accuracy(%)	Weighted F1-score	Accuracy(%)	Weighted F1-score	Accuracy(%)	Weighted F1-score
Average	91.74	0.90	88.50	0.87	81.71	0.79
HAPV(e.g., happy, excited)	**92.89**	**0.92**	**89.83**	0.88	**83.33**	**0.83**
HANV(e.g., fear, anger)	91.56	0.91	88.33	**0.89**	81.17	0.79
LANV(e.g., bored, sad)	89.17	0.89	87.91	0.86	78.23	0.78
LAPV(e.g., relief, serene)	90.78	0.88	87.50	0.87	81.78	0.80

Note: We bold the highest values while underlining the lowest.

TABLE 7
Comparison of EmoTake performance when removing a certain channel.

Backbone	Channels	Takeover readiness		Takeover reaction time		Takeover quality	
		Accuracy(%)	Weighted F1-score	Accuracy(%)	Weighted F1-score	Accuracy(%)	Weighted F1-score
iTransformer	Action Units(-)	**87.17**	**0.87**	**81.53**	**0.81**	**77.17**	**0.77**
	Eye Movement(-)	87.89	0.87	85.78	0.85	78.12	0.78
	Head Pose(-)	90.44	0.89	85.05	0.84	80.33	0.79
	Body Posture(-)	89.67	0.89	82.91	0.82	78.23	0.78
	All	91.74	0.90	88.50	0.87	81.71	0.79

(-) represents removal for the mentioned factor.

(DNN) [46], along with the Long Short-Term Memory (LSTM) [6] specialized for processing time-series data. Additionally, traditional machine learning classifiers such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) were also included in the comparison, with 10 manual features, i.e., the median, standard deviation, variance, minimum, maximum, Q1 (minimum quartile), Q3 (maximum quartile), range, skewness, and kurtosis. The results are concluded in Fig. 8. In addition, due to the imbalance of the sample labels, we performed the prediction results on the majority class in these figures.

Firstly, we discovered that iTransformer exhibits promising performance in dealing with multi-modal time series data against DNN and LSTM. This is because it can learn variate-centric representations with meaningful attention maps across multi-modal time series data by inverting the duties of the attention mechanism and the feed-forward network. Secondly, we found that traditional classifiers like SVM, KNN, and RF also work well on this problem. We think that this is because instead of processing images directly, we have quantified drivers' physical and emotional states into time-series numerical data, which can be properly processed by these classic classifiers. In addition, when analyzing the prediction results of the majority class, we found that the prediction results of the majority class were similar to the results of the full-volume analysis on takeover readiness and reaction time. This indicates that our model adequately learned the features of the different classes rather than incorrectly predicting the minority class as the majority class. The majority class shows higher accuracy and F1-score on takeover quality, we believe that it is led by the more detailed classes of takeover quality, which causes the difficulty of identification.

6 Discussion

6.1 Rate Agreement

We hired four raters to label the takeover quality. The raters all have at least a bachelor's degree and two years of driving experience. Specific information is given in Table 8.

TABLE 8
Overview on the raters (A-D) by age, gender, degree, and driving experience.

Rater	Age(years)	Gender	Degree	Driving exp.(years)
A	23	m	bachelor	4
B	25	f	master	3
C	24	f	bachelor	2
D (senior)	37	m	Ph.D	9

The two-round annotation is done independently by each rater in a quiet environment. The first round is during the experiment, where three raters conducted preliminary scoring based on the real-time performance of the experimenter during the takeovers, and the scoring is based on ISO/TR $21959-1$ (Road vehicles — Human performance and state in the context of automated driving), with a rating scale of 1-5 (where 1 is the worst and 5 is the best). The second round is after all experiments have been completed, where one senior rater reviews the driving data collected during the experiment to refine the preliminary scores in the first round.

TABLE 9
Overview on the raters' (ID A-D) agreement: *Pearson*, *CCC*, *kappa* and *weighted kappa* of the individual raters and EWE(mean).

ID	*Pearson*	*CCC*	*kappa*	*weighted kappa*
A	0.913	0.907	0.796	0.851
B	0.944	0.940	0.874	0.906
C	0.942	0.937	0.862	0.899
D	0.952	0.950	0.894	0.921

During the process, raters took a 10-minute break every 30 minutes to avoid fatigue annotation. After completing the annotation, we establish a gold standard for these four subjective scores via leveraging the Evaluator Weighted Estimator (EWE) [62] fusion for the average value. We also perform a simple rounding to maintain the original five categories. Subsequently, we compare each rater's score

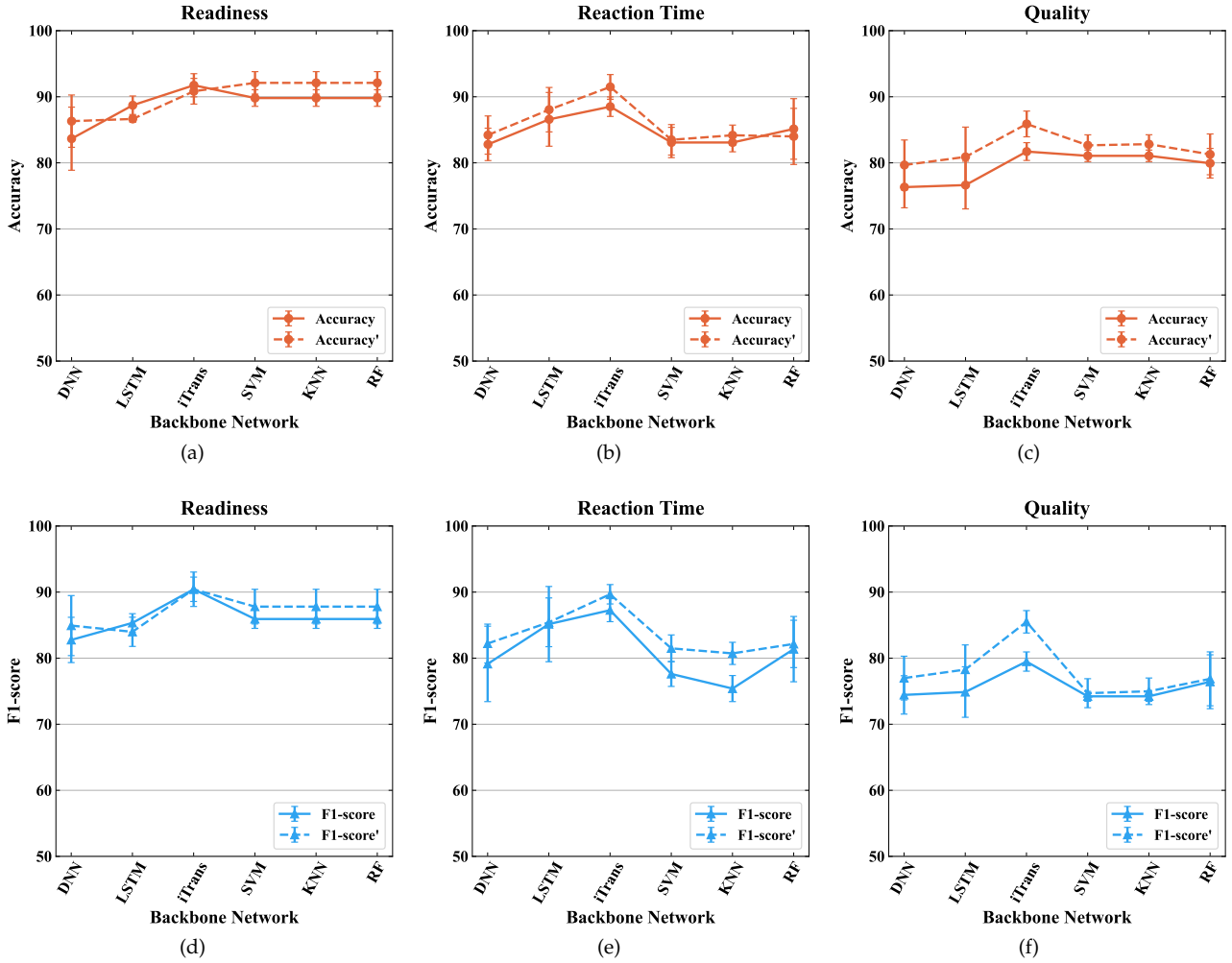


Fig. 8. Naive comparison under subject-independent cross-validation. (DNN: Deep Neural Networks [46]; LSTM: Long Short-Term Memory [6]; iTrans: inverted Transformer; SVM: Support Vector Machine; KNN: K-Nearest Neighbor; RF: Random Forest); Accuracy': Accuracy of majority class; F1-score': F1-score of majority class;)

with the gold standard, and analyze the Pearson correlation coefficient (Pearson), consistency correlation coefficient (CCC), Kappa correlation coefficient, and weighted Kappa correlation coefficient between them, as shown in Table 9. The results show a high level of consistency among raters, indicating the effectiveness of the labels of samples.

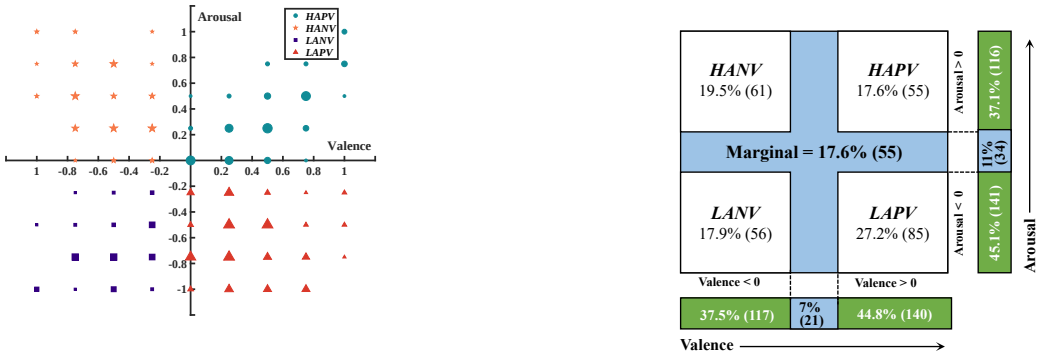
6.2 Emotion Induction

As shown in Fig. 9 (a), our emotion induction method has successfully stimulated diverse emotions covering all four quadrants on the valence-arousal plane. Specifically, 17.6%, 19.5%, 17.9% and 27.2% of takeover records fall into ***HAPV***, ***HANV***, ***LANV*** and ***LAPV***, respectively. Please note that as shown in Fig. 9 (b), we have discarded the region in blue for analysis where arousal or valence equals 0 to reduce category ambiguity.

As reported in Section 4.3.2. the first quadrant (high arousal and positive valance, ***HAPV***) contains 32 video clips lasting for 85.1 minutes in total. The usual emotions in ***HAPV*** include happiness and excitement, which are relatively harder to be induced under experimental pressure. Therefore, it has the most clips and the shortest length per clip (2.66 minutes) so that the participants have enough choices to reach the mood. They described their feelings as excited, arousing, and pleasant. For example, *P*1 said: "I found some of these videos were so funny...my face hurt from laughing." On the contrary, emotions in the third quadrant (low arousal and negative valence, ***LANV***) like bored and sad are much easier to stimulate as participants reported that a long and boring video made them feel sleepy. As a result, ***LANV*** has the longest clips on average, i.e., 6.3 minutes per clip. Participants described their feelings as bored and listless. For example, *P*3 said: "*So bored...I almost fell asleep.*" and *P*20 said: "*I could not stop blinking.*"

Typical emotions in the second quadrant (high arousal and negative valence, ***HANV***) include fear and anger, which exhibit significant gender differences. For example, *P*6 (female) said: "*I was so afraid of the horror clips. I need to pray for no nightmare tonight.*", while *P*9 (male) said for the same clips: "*Just so so, compared to what I have seen before.*"

The induction clips are much gentler in the fourth quadrant (low arousal and positive valence, ***LAPV***) to keep the participants in a relief and soothing mood. They described feeling as peaceful, comfortable, and enjoyable. For example, *P*12 said: "*It was nice in automatic driving so I can enjoy*



(a) (b)

Fig. 9. (a) Distribution of induced self-rated emotional states in the Valence-Arousal plane. A larger marker size represents a larger quantity. (b) Counts of self-rated emotional states in different quadrants. The region in blue (where arousal or valence equals 0) is discarded for analysis to reduce category ambiguity.

the beautiful view.". Interestingly, some participants reported being aroused in these spectacular natural scenarios. $P5$ said: *"The scene was so beautiful and made me excited."*

The audio-visual stimuli are proven to be effective in our experiments. However, considering our setting, we assume the drivers are watching video clips as NDRTs. But in the real world, we would like to explore side channel stimuli like scents [15] for emotion regulation, since they pose fewer safety risks.

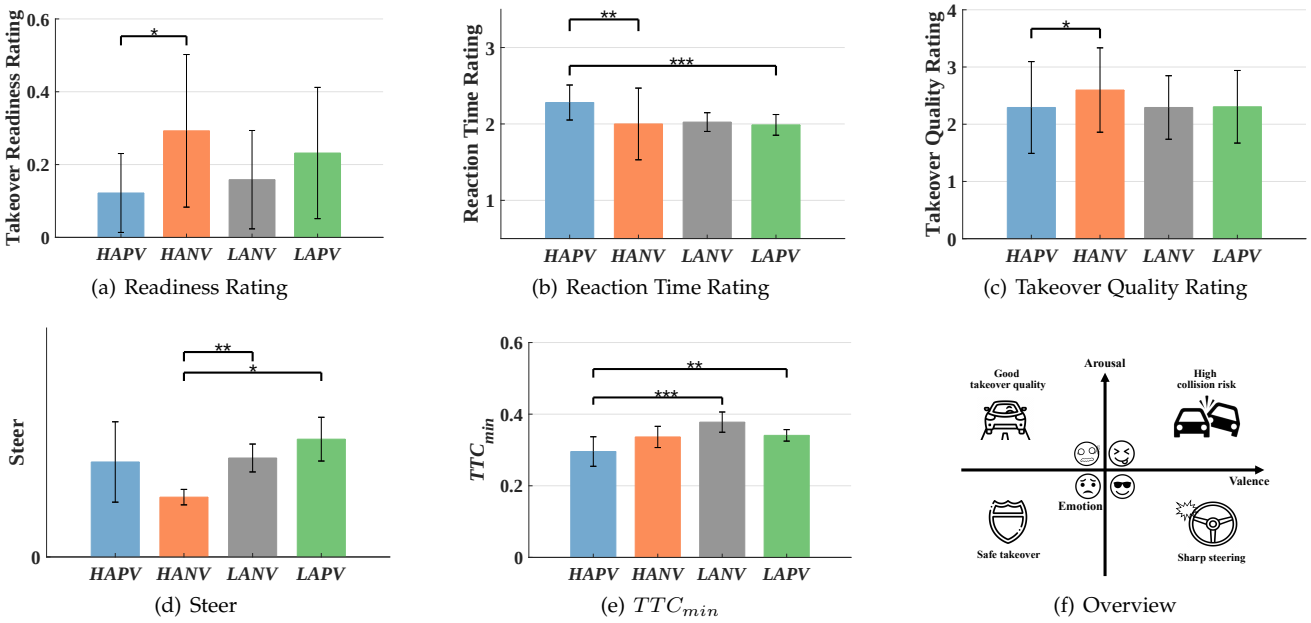
6.3 Emotion and Takeover Behaviors

To understand how diverse emotions affect takeover behavior, we first performed a correlation analysis between valence/arousal and the dependent variables of takeover behaviors (see Table 10). Then we studied the relationship between emotional states in four quadrants with these dependent variables. Data from one participant was excluded from the analysis as the participant did not follow the instructions from the experimenter.

6.3.1 Valence and Takeover Behavior

We collected 300 takeover events of participants, as well as their self-reported emotional state ratings. On a scale of 1 to 9 (obtained from the SAM questionnaire), valence rates had an average value of 5.14 and an STD of 2.16. A Spearman test was employed to test the monotonic correlations between valence rates and takeover behaviors.

Based on our results, valence was negatively correlated with TTC_{min} ($r = -0.16, p < 0.005$). Additionally, a paired t-test showed that the TTC_{min} of positive valence was significantly shorter than negative valence ($t = -2.3, p < 0.03$). The difference may partly be due to that low arousal is beneficial to driving safety. However, no other significant effect was found.



(a) Readiness Rating (b) Reaction Time Rating (c) Takeover Quality Rating

(d) Steer (e) TTC_{min} (f) Overview

Fig. 10. The mean values of takeover readiness rating (0 = "Not Ready", 1 = "Ready"), reaction time rating (1 = "Long", 3 = "Short"), quality rating (1 = "Worst", 5 = "Perfect") with the emotional states of ***HAPV***, ***HANV***, ***LAPV***, and ***LANV*** (a-c); the mean values of normalized steering angle and normalized TTC_{min} (d-e). Error bars, ± s.e.m., $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

TABLE 10
Dependent variables of takeover behaviors

Features	Unit	Explanation
Readiness	Binary	How ready the participant were to take control of the vehicle
Reaction time	Scale from 1-3	Three levels according to reaction time
Quality	Scale from 1-5	The controllability of transitions that aggregates different aspects of the takeover situation to one global measure of driving quality
Vehicle data		
Speed	m/s	Driving speed during obstacle avoidance
Accelerate	m/s^2	Acceleration during obstacle avoidance
Jerk	m/s^3	The derivative of acceleration during obstacle avoidance
Brake	degree	The angle of brakes during obstacle avoidance
Steer	degree	Steering wheel rotation angle during obstacle avoidance
TTC_{min}	ms	Minimum time to collision

6.3.2 Arousal and Takeover Behavior

The arousal rates had an average value of 4.35 and an STD of 2.32. We applied the same correlation tests to arousal. Based on our results, arousal is negatively correlated with steer angle ($r = -0.21, p < 0.001$) and TTC_{min} ($r = -0.14, p < 0.02$). Additionally, a paired t-test showed that the TTC_{min} of high arousal is significantly shorter than low arousal ($t = -2.4, p < 0.02$). The difference may partly be due to the higher immersion of NDRT in high arousal leading to higher collision risk. However, no significant effect was found between the steering angle of high arousal and low arousal.

6.3.3 Emotional States and Takeover Behavior

As mentioned in the previous section, we divided the emotional state rating into four quadrants based on their valence and arousal rates (***HAPV***, ***HANV***, ***LAPV***, and ***LANV***). In this part of our analysis, we used these labels as emotional states and evaluated their impact on takeover behaviors. The dependent variables of takeover behaviors are described in Table 10. Since the Shapiro-Wilk test showed a significant departure from normality for all the variables of takeover behaviors, we compared the means of variables in the four emotional states using the non-parametric Kruskal-Wallis Test. We leverage the Wilcoxon rank sum test with False Discovery rate (FDR) correction to get the significance level between quadrants due to multiple comparisons.

On one hand, we find statistically significant differences exist in all three takeover behavior ratings, i.e., takeover readiness ($\chi^2(3) = 8.85, p = 0.031$), reaction time (1 = "Long", 3 = "Short") ($\chi^2(3) = 18.91, p = 0.0003$) and takeover quality (1 = "Worst performance", 5 = "Perfect performance") ($\chi^2(3) = 6.24, p = 0.041$). On the other hand, we find that the impact of different emotions on different takeover behaviors is not always consistent. For takeover readiness and takeover quality, the post hoc pairwise analysis points out the difference between *HAPV* and *HANV* ($p_{fdr} = 0.0212$ and $p_{fdr} = 0.0447$, respectively). But such significant differences happen not only between *HAPV* and *HANV* ($p_{fdr} = 0.0041$) but also between *HAPV* and *LAPV* ($p_{fdr} = 0.0003$) as well for reaction time.

We further analyze the vehicle data (e.g., speed, acceleration, etc.) between four emotional states to understand how emotions affect takeover quality more specifically. The Kruskal Wallis test showed a significant difference between the steering angle of four emotional states ($\chi^2(3) = 12.63, p = 0.013$). The analysis showed differences between *HANV* and *LAPV* ($p_{fdr} = 0.0116$) and significant differences between *HANV* and *LANV* ($p_{fdr} = 0.0048$). Moreover, there is a significant difference in the minimum time to collision TTC_{min} between emotional states ($\chi^2(3) = 17.69, p = 0.0005$). The analysis showed differences between *HAPV* and *LAPV* ($p_{fdr} = 0.0061$) and significant difference between *HAPV* and *LANV* ($p_{fdr} = 0.0001$). However, no significant differences were found in other objective takeover measures.

In summary, as shown in Fig. 10, drivers in ***HAPV*** had a lower takeover readiness, quicker reaction, lower takeover quality, and smaller TTC_{min}, indicating a higher risk of collision. Drivers in ***HANV*** had a higher takeover readiness, lower reaction score, higher takeover quality, and a smaller angle of steering, leading to a better takeover quality. Drivers in ***LANV*** had a larger angle of steering, and a larger TTC_{min}, suggesting a safer takeover. Lastly, drivers in ***LAPV*** had a lower reaction score and the largest angle of steering, resulting in a radical change of direction.

7 SUGGESTIONS

Through performance evaluation and the discussion, we have the following suggestions for designers who are interested in exploring emotions in vehicular applications like takeover prediction and emotion regulation,

- Considering vision as an effective signal to monitor drivers' states since the cutting-edge deep neural networks enable the collection of drivers' multi-channel emotional and physical data.
- Valuing facial expression in predicting driver takeover behaviors since it is an important channel for expressing emotions that affect their behaviors.
- High arousal possibly leads to intensive body movements, which affect the drivers' both emotional and physical states, making it easier to be recognized by the network.
- Paying more attention to emotions in ***HAPV*** (e.g., happy and exciting), since drivers in this quadrant are more immersed in NDRTs and thus have higher collision risks. Regulating drivers' emotions from

HAPV to ***LAPV*** (lower collision risk) via a proper induction like the video clips we collected (e.g., natural sceneries) could greatly improve takeover safety.

- Regulating drivers' emotions to a target quadrant in autopilot mode is possible, and sometimes very easy to do so if given a proper way (e.g., audiovisual stimuli).

8 Limitation

This work has several limitations. First, our user study is limited to a driving simulator and primarily involves college students. We cannot guarantee the applicability of our model and results in real driving conditions or other populations. Second, the study was conducted during the daytime, potentially affecting performance in nighttime conditions due to increased noise in camera data. Possible remedies include enhancing the framework with data pre-processing techniques and utilizing deep networks designed for noisy data or leveraging a camera with infrared night vision, which is slighter expensive than the one we used (e.g., HIKVISION 3367WDV3, 4K, $ 77). Third, we chose video watching as the NDRT because it is one of the most common activities in driving and it could effectively stimulate drivers' diverse emotions. How our model performs in other common NDRTs like reading, working, and talking to others (via mobile or to other passengers) remains unknown. Fourth, facial expression is an important channel in our model for prediction. Some may worry about the privacy issue. A possible solution is to use a depth camera that measures only the depth information of the face (especially the eye and mouth regions) [63] instead of the RGB camera we used. Lastly, a single camera alone may not be the right setup as a highly critical, possibly life-saving driving assistance. Its capacity is limited by various conditions like illumination, sunglasses, mask, and head pose.

9 Conclusion

In this work, we presented the first system that explicitly decodes drivers' facial expressions for predicting their takeover behaviors via a unified deep neural framework. Analysis of results from a user study with 26 participants provided us with more insights into emotion induction, emotion in prediction, and how emotions affect takeover behaviors. Our work suggests that facial expression constitutes an important channel in predicting takeover behaviors. Also, different emotions have diverse influences on takeover behaviors, and certain emotional states like positive valence deserve more attention because they imply a higher collision risk. As for the next step, it is our priority to collaborate with local automobile manufacturers and seek to conduct real-world tests in the future to clarify a series of unresolved issues, including the limitations of cameras.

References

[1] P. Bazilinskyy and J. de Winter, "Auditory interfaces in automated driving: an international survey," *PeerJ Computer Science*, vol. 1, p. e13, 2015.

[2] N. Kim, K. Jeong, M. Yang, Y. Oh, and J. Kim, "" are you ready to take-over?" an exploratory study on visual assistance to enhance driver vigilance," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 1771–1778.

[3] F. Meng, C. Ho, R. Gray, and C. Spence, "Dynamic vibrotactile warning signals for frontal collision avoidance: towards the torso versus towards the head," *Ergonomics*, vol. 58, no. 3, pp. 411–425, 2015.

[4] E. Pakdamanian, S. Sheng, S. Baee, S. Heo, S. Kraus, and L. Feng, "Deeptake: Prediction of driver takeover behavior using multimodal data," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[5] J. Ayoub, N. Du, X. J. Yang, and F. Zhou, "Predicting driver takeover time in conditionally automated driving," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[6] N. Deo and M. M. Trivedi, "Looking at the driver/rider in autonomous vehicles to predict take-over readiness," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 1, pp. 41–52, 2019.

[7] N. Du, F. Zhou, E. M. Pulver, D. M. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang, "Predicting driver takeover performance in conditionally automated driving," *Accident Analysis & Prevention*, vol. 148, p. 105748, 2020.

[8] C. Braunagel, W. Rosenstiel, and E. Kasneci, "Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 10–22, 2017.

[9] M. Jeon, "Emotions and affect in human factors and human–computer interaction: Taxonomy, theories, approaches, and methods," *Emotions and affect in human factors and human-computer interaction*, pp. 3–26, 2017.

[10] N. Du, F. Zhou, E. M. Pulver, D. M. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang, "Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving," *Transportation research part C: emerging technologies*, vol. 112, pp. 78–87, 2020.

[11] A. Eherenfreund-Hager, O. Taubman-Ben-Ari, T. Toledo, and H. Farah, "The effect of positive and negative emotions on young drivers: A simulator study," *Transportation research part F: traffic psychology and behaviour*, vol. 49, pp. 236–243, 2017.

[12] H. K. Sanghavi, "Exploring the influence of anger on takeover performance in semi-automated vehicles," Ph.D. dissertation, Virginia Tech, 2020.

[13] J. Schmidt, M. Dreißig, W. Stolzmann, and M. Rötting, "The influence of prolonged conditionally automated driving on the take-over ability of the driver," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2017, pp. 1974–1978.

[14] "Road vehicles - human performance and state in the context of automated driving." [Online]. Available: https://doi.org/10.3403%2F30362050u

[15] D. Dmitrenko, E. Maggioni, G. Brianza, B. E. Holthausen, B. N. Walker, and M. Obrist, "Caroma therapy: Pleasant scents promote safer driving, better mood, and improved well-being in angry drivers," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–13. [Online]. Available: https://doi.org/10.1145/3313831.3376176

[16] S. Zepf, M. Dittrich, J. Hernandez, and A. Schmitt, "Towards empathetic car interfaces: Emotional triggers while driving," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–6. [Online]. Available: https://doi.org/10.1145/3290607.3312883

[17] L. Mou, Y. Zhao, C. Zhou, B. Nakisa, M. N. Rastgoo, L. Ma, T. Huang, B. Yin, R. Jain, and W. Gao, "Driver emotion recognition with a hybrid attentional multimodal fusion framework," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2970–2981, 2023.

[18] L. F. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition & Emotion*, vol. 12, no. 4, pp. 579–599, 1998.

[19] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[20] R. Abdu, D. Shinar, and N. Meiran, "Situational (state) anger and driving," *Transportation research part F: traffic psychology and behaviour*, vol. 15, no. 5, pp. 575–580, 2012.

[21] H. Hu, Z. Zhu, Z. Gao, and R. Zheng, "Analysis on biosignal characteristics to evaluate road rage of younger drivers: A driving simulator study," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 156–161.

[22] M. Jeon, B. N. Walker, and J.-B. Yim, "Effects of specific emotions on subjective judgment, driving performance, and perceived workload," *Transportation research part F: traffic psychology and behaviour*, vol. 24, pp. 197–209, 2014.

[23] A. Safety, "Prevalence of self-reported aggressive driving behavior: United states, 2014. (july 2016)."

[24] T. A. Dingus, F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2636–2641, 2016.

[25] T. Zimasa, S. Jamson, and B. Henson, "Are happy drivers safer drivers? evidence from hazard response times and eye tracking data," *Transportation research part F: traffic psychology and behaviour*, vol. 46, pp. 14–23, 2017.

[26] J. Lu, X. Xie, and R. Zhang, "Focusing on appraisals: How and why anger and fear influence driving risk perception," *Journal of safety research*, vol. 45, pp. 65–73, 2013.

[27] Y.-C. Liu and T.-J. Wu, "Fatigued driver's driving behavior and cognitive task performance: Effects of road environments and road environment changes," *Safety Science*, vol. 47, no. 8, pp. 1083–1089, 2009.

[28] M. Chan and A. Singhal, "The emotional side of cognitive distraction: Implications for road safety," *Accident Analysis & Prevention*, vol. 50, pp. 147–154, 2013.

[29] G. Hancock, P. Hancock, and C. Janelle, "The impact of emotions and predominant emotion regulation technique on driving performance," *Work*, vol. 41, no. Supplement 1, pp. 3608–3611, 2012.

[30] C. Pêcher, C. Lemercier, and J.-M. Cellier, "Emotions drive attention: Effects on driver's behaviour," *Safety Science*, vol. 47, no. 9, pp. 1254–1259, 2009.

[31] J. Navarro, F. Osiurak, and E. Reynaud, "Does the tempo of music impact human behavior behind the wheel?" *Human factors*, vol. 60, no. 4, pp. 556–574, 2018.

[32] L. M. Trick, S. Brandigampola, and J. T. Enns, "How fleeting emotions affect hazard perception and steering while driving: The impact of image arousal and valence," *Accident Analysis & Prevention*, vol. 45, pp. 222–229, 2012.

[33] A. Lotz and S. Weissenberger, "Predicting take-over times of truck drivers in conditional autonomous driving," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2018, pp. 329–338.

[34] B. Zhang, J. de Winter, S. Varotto, R. Happee, and M. Martens, "Determinants of take-over time from automated driving: A meta-analysis of 129 studies," *Transportation research part F: traffic psychology and behaviour*, vol. 64, pp. 285–307, 2019.

[35] C. Gold, M. Körber, D. Lechner, and K. Bengler, "Taking over control from highly automated vehicles in complex traffic situations: the role of traffic density," *Human factors*, vol. 58, no. 4, pp. 642–652, 2016.

[36] N. Du, J. Kim, F. Zhou, E. Pulver, D. M. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang, "Evaluating effects of cognitive load, takeover request lead time, and traffic density on drivers' takeover performance in conditionally automated driving," in *12th international conference on automotive user interfaces and interactive vehicular applications*, 2020, pp. 66–73.

[37] S. C. Lee, S. H. Yoon, and Y. G. Ji, "Effects of non-driving-related task attributes on takeover quality in automated vehicles," *International Journal of Human–Computer Interaction*, vol. 37, no. 3, pp. 211–219, 2021.

[38] K. Zeeb, M. Härtel, A. Buchner, and M. Schrauf, "Why is steering not the same as braking? the impact of non-driving related tasks on lateral and longitudinal driver interventions during conditionally automated driving," *Transportation research part F: traffic psychology and behaviour*, vol. 50, pp. 65–79, 2017.

[39] B. Wandtner, N. Schömig, and G. Schmidt, "Effects of non-driving related task modalities on takeover performance in highly automated driving," *Human factors*, vol. 60, no. 6, pp. 870–881, 2018.

[40] J. J. So, S. Park, J. Kim, J. Park, and I. Yun, "Investigating the impacts of road traffic conditions and driver's characteristics on automated vehicle takeover time and quality using a driving simulator," *Journal of advanced transportation*, vol. 2021, 2021.

[41] H. Clark and J. Feng, "Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation," *Accident Analysis & Prevention*, vol. 106, pp. 468–479, 2017.

[42] Y.-K. Ou, W.-X. Huang, and C.-W. Fang, "Effects of different takeover request interfaces on takeover behavior and performance during conditionally automated driving," *Accident Analysis & Prevention*, vol. 162, p. 106425, 2021.

[43] A. L. Kun, S. Boll, and A. Schmidt, "Shifting gears: User interfaces in the age of autonomous driving," *IEEE Pervasive Computing*, vol. 15, no. 1, pp. 32–38, 2016.

[44] N. Du, F. Zhou, E. Pulver, D. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang, "Predicting takeover performance in conditionally automated driving," in *Extended abstracts of the 2020 chi conference on human factors in computing systems*, 2020, pp. 1–8.

[45] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–30, 2020.

[46] Y. Wang, Y. Gu, and F. Ren, "Emotion-aware takeover performance prediction system in semi-autonomous driving," *IEEE Communications Magazine*, vol. 61, no. 10, pp. 70–75, 2023.

[47] S. K. D'Mello, S. D. Craig, and A. C. Graesser, "Multimethod assessment of affective experience and expression during deep learning," *International Journal of Learning Technology*, vol. 4, no. 3-4, pp. 165–187, 2009.

[48] A. Kapoor, S. Mota, R. W. Picard *et al.*, "Towards a learning companion that recognizes affect," in *AAAI Fall symposium*, vol. 543, 2001, pp. 2–4.

[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[50] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[51] S. Ansari, H. Du, F. Naghdy, and D. Stirling, "Automatic driver cognitive fatigue detection based on upper body posture variations," *Expert Systems with Applications*, p. 117568, 2022.

[52] D. Osokin, "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose," *arXiv preprint arXiv:1811.12004*, 2018.

[53] D. Crundall and G. Underwood, "Visual attention while driving: measures of eye movements used in driving research," in *Handbook of traffic psychology*. Elsevier, 2011, pp. 137–148.

[54] Y. Li, J. Li, X. Jiang, C. Gao, and T. Zhang, "A driving attention detection method based on head pose," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2019, pp. 483–490.

[55] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[56] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.

[57] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[58] H. Wu, C. Wu, N. Lyu, and J. Li, "Does a faster takeover necessarily mean it is better? a study on the influence of urgency and takeover-request lead time on takeover performance and safety," *Accident Analysis & Prevention*, vol. 171, p. 106647, 2022.

[59] K. S. Quigley, K. A. Lindquist, and L. F. Barrett, "Inducing and measuring emotion and affect: Tips, tricks, and secrets." 2014.

[60] W. Li, Y. Cui, Y. Ma, X. Chen, G. Li, G. Zeng, G. Guo, and D. Cao, "A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios," *IEEE Transactions on Affective Computing*, 2021.

[61] S. Ghosh, G. Pons Rodriguez, S. Rao, A. El Ali, and P. Cesar, "Exploring emotion responses toward pedestrian crossing actions for designing in-vehicle empathic interfaces," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–6.

[62] B. W. Schuller, *Intelligent audio analysis*. Springer, 2013, vol. 3.

[63] X. Kong, Z. Meng, L. Meng, and H. Tomiyama, "A privacy protected fall detection iot system for elderly persons using depth camera," in *2018 International Conference on Advanced Mechatronic Systems (ICAMechS)*. IEEE, 2018, pp. 31–35.