Efficient Mixture-of-Expert for Video-based Driver State and Physiological Multi-task Estimation in Conditional Autonomous Driving

Jiyao Wang, *Student Member, IEEE,* Xiao Yang, Zhenyu Wang, Ximeng Wei, Ange Wang, Dengbo He, Kaishun Wu *Fellow, IEEE*

***Abstract*—Road safety remains a critical challenge worldwide, with approximately 1.35 million fatalities annually attributed to traffic accidents, often due to human error. As we advance towards higher levels of vehicle automation, challenges still exist, as driving with automation can cognitively over-demand drivers if engaging in non-driving-related tasks (NDRTs), or lead to drowsiness if driving is the sole task. This calls for the urgent need for an effective Driver Monitoring System (DMS) that can evaluate cognitive load and drowsiness in SAE Level-2/3 autonomous driving contexts. In this study, we propose a novel multi-task DMS, termed VDMoE, which leverages RGB video input to monitor driver states non-invasively. By utilizing key facial features to minimize computational load and integrating remote Photoplethysmography (rPPG) for physiological insights, our approach enhances detection accuracy while maintaining efficiency. Additionally, we optimize the Mixture-of-Experts (MoE) framework to accommodate multi-modal inputs and improve performance across different tasks. A novel prior-inclusive regularization method is introduced to align model outputs with statistical priors, thus accelerating convergence and mitigating overfitting risks. We validated our method by creating a new dataset (MCDD), which comprises RGB video and physiological indicators from 42 participants, as well as two public datasets. Our findings demonstrate the effectiveness of VDMoE in monitoring driver states, contributing to safer autonomous driving systems. The code and data will be released.**

***Index Terms*—Driver Monitoring System, cognitive load, drowsiness detection, multi-task learning, Mixture-of-Experts.**

I. Introduction

ROAD safety is still an extreme challenge for societies [1], [2]. The World Health Organization (WHO) reported that approximately 1.35 million people die each year as a result of road traffic accidents, with human error being a significant contributing factor. While autonomous technologies promise to reduce human error—the leading cause of traffic accidents—there remains a critical period where human drivers must share control with automated systems. Particularly, at the Society of Automotive Engineers (SAE) Level-2 (L2) automation [3], the vehicle can control both steering and acceleration/deceleration but still requires the driver to remain actively engaged and ready to take over at any moment; while with SAE Level-3 (L3) automation, the vehicle can handle most driving tasks but still requires human drivers to step in when prompted by the take-over requests (TORs). Both have shifted the role of the driver from an operator to a supervisor. As a result, drivers are more likely to experience low cognitive load if driving is the sole task, or more likely to be cognitively over-demanding, if drivers are engaged in non-driving-related tasks (NDRTs), both can be detrimental to driving performance [4], [5]. For example, the cognitive overload induced by multi-dimensional cognitive resource consumption can lead to a limited field of vision [6], and a reduced ability to foresee potential hazards [7]. Further, the effect of long-time driving drowsiness and the high cognitive load as a result of NDRTs may co-exist, leading to impaired driving performance [8], [9]. Thus, the Driver Monitoring System (DMS) that is capable of estimating drivers' cognitive load and drowsiness is critical to enhancing the safety and reliability of L2/L3 vehicles.

DMS has been widely studied in recent years. Traditional DMS relied on various sensors, including physiological sensors (e.g., ECG sensors on the steering wheel) and vehicle-based sensors (e.g., steering angle, lane departure warnings) [10]. However, in SAE L2 or L3 vehicles, the driver's manual driving performance data is not available most of the time, as the vehicle is controlled by the driving automation. Physiological signals, though can be served as effective indicators of driver's state [11], [12], usually depend on invasive physiological sensors (e.g., electroencephalography (EEG) [13], [14], electrooculography (EOG), [15]) and thus are currently not practically applicable in vehicle cabins. Although there were some attempts to leverage various non-invasive physiological sensors [16], [17], drivers were still required to wear inconvenient and costly signal acquisition equipment, which is still far from large-scale commercialization. Thus, considering the feasibility and cost of real-world in-vehicle deployment, camera is still a compelling alternative to monitor driver states non-intrusively [18]. Especially, the fast development in deep computer vision enables efficient non-contact state detection, and thus, in recent years, several facial image-based driver drowsiness detection methods have been proposed [19], [20].

Manuscript created October 2025; This work was supported by the National Natural Science Foundation of China (No. 52202425), Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), and Guangzhou Science and Technology Program City-University Joint Funding Project (No. 2023A03J0001) (Corresponding author: Dengbo He).

Jiyao Wang, Xiao Yang, Zhenyu Wang, Ange Wang, and Dengbo He is with the Systems Hub, the Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, (e-mail: jwanggo@connect.ust.hk; xyang856@connect.hkust-gz.edu.cn; zwang209@connect.hkust-gz.edu.cn; awang324@connect.hkustgz.edu.cn; dengbohe@hkust-gz.edu.cn). Ximeng Wei is with the University of Hong Kong, Hong Kong SAR, China, (e-mail: aasimon0827@gmail.com). Kaishun Wu is with the Information Hub, the Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, (e-mail: wuks@hkust-gz.edu.cn).

TABLE I
COMPARISON OF PUBLIC NON-SYNTHETIC DATASET FOR STATE ESTIMATION.

Dataset	Date	Subject	Camera Setting	HR	RR	Drowsiness	cognitive load	Application Scenario
DROZY [29]	2016	14	NIR	✓	✗	✓	✗	-
NTHU [30]	2016	36	RGB, IR	✗	✗	✓	✗	Driving
UTA-RLDD [31]	2019	60	RGB	✗	✗	✓	✗	-
DMD [32]	2020	37	RGB, IR, Depth	✗	✗	✓	✗	Driving
FatigueView [33]	2022	95	RGB, IR	✗	✗	✓	✗	Driving
MOCAS [34]	2024	21	RGB	✓	✗	✗	✓	-
MCDD	2024	42	RGB	✓	✓	✓	✓	Driving

Notes: NIR means the Near Infrared camera, IR means the Infrared camera.

However, most of the previous methods were based on single-frame detection, which loses temporal variation in both surface facial expression and internal physiological reflection. Compared to image-based methods, video-based methods [21], [18], [22] can make effective use of continuous temporal features and have the potential to enhance detection accuracy. However, existing methods still have the following shortcomings: (1) some video-based drowsiness detection methods [23] directly utilize continuous frames as the input, leading to significant computational resources and interference from redundant information; (2) some methods [18], [24] tried to replace video by key optimized facial features (e.g., landmarks, eye and mouth areas), while they cannot assess physiological features without supplementing signals from physiological sensors; (3) current video-based DMS focused mostly on single-task drowsiness [21], [18], [22] or distraction [25] estimation. However, driver states do not appear independently; instead, the interaction effects among multiple states make driver state estimation complicated. For example, in the context of conditional autonomous driving (i.e., driving with SAE L2/L3 vehicles), the cognitive load can moderate both drowsiness [26] and physiological reflection [27]. However, few non-contact solutions were proposed for cognitive load estimation, drowsiness detection and physiological reflections. Given that training separate models for each task incurs high deployment costs and reduces iterative efficiency [28], it is urgent to develop a unified model and make full use of the association across drivers' multiple states.

To address the aforementioned issues, we propose an RGB video-based multi-task DMS (VDMoE). Specifically, based on previous studies [18], [35], we first utilized the key facial features to reduce redundant information and computational costs. At the same time, considering cardiac and respiration activities are crucial predictors of driver states [27], [26], and are also important driver health indicators, we integrated the remote Photoplethysmography (rPPG) technology [36] into our model. Since the facial landmark and local eye and mouth region process contain limited temporal color change of the facial skin pixel points, which can be converted to a change of the blood volume under the skin [28], we first selected key facial regions of interest based on facial landmarks. Then, by transforming the color space (from RGB to YUV) as well as by band-pass denoising, we obtained the alternative multi-modal information input Spatial-temporal Map (STMap). Next, to fully leverage the dependencies between different tasks, the Mixture-of-Experts (MoE) [37] structure is introduced. Being different from the classic MoE structure, the heterogeneous gating mechanism and spatio-temporal expert separation are designed for the multi-modal alternative inputs to improve the multi-task performance. Besides, recent video-based DMS used various types of neural networks (e.g., CNN [21], Transformer [18]) that are suitable for processing different types of data or extracting different levels of feature information [38]. However, given the demand for real-time assessment of DMS, networks with excessive parameters and computational complexity will pose a challenge to deployment hardware. Therefore, we instantiated the nonlinear feature learning component of the basic block of VDMoE as a simple two-layer multi-layered perception (MLP) network to replace the previous modules represented by residual convolution [39] or multi-head attention [40]. Lastly, for the optimization goal, to explicitly capture the dependency between different states, we strengthen the learning capacity of VDMoE through one regularization based on prior knowledge from human factors fields [26]. Given there is no proper dataset for multi-task video-based DMS development that considered both the cognitive load and drowsiness in the driving automation context (seeing Table I), a driving simulator experiment with 42 participants was conducted. Extensive experiments based on our dataset illustrate the effectiveness of our method. In all, the contributions of our work are summarized as follows:

- A multi-task RGB video-based driver state monitoring method (VDMoE) is proposed in this work. This system concentrates on the drowsiness and physiological reflection moderated by the multi-dimensional cognitive load in the context of driving automation. As far as we know, this is the first multi-task model that can simultaneously estimate driver drowsiness, cognitive load, and cardiac and respiratory activities.
- To achieve the balance between estimation accuracy and efficiency, we replaced the full video with key facial features (landmarks, eye, and mouth area) as input, and introduced the STMap representing periodical color changes in facial regions to supplement alternative physiological features.
- Inspired by the superior performance of MoE in multi-task learning, we further optimized the MoE structure

to accommodate the multi-modal and spatio-temporal input through a heterogeneous gating mechanism and spatio-temporal expert separation. At the same time, we replaced more complex network structures with MLPs for lightweight and future deployment. Comparison of computational cost with existing baselines proves our proposals' effectiveness.
- A novel prior-inclusive regularization, inspired by human factor engineering studies, was proposed to enforce the multi-task model's outputted probability distributions for each state to be aligned with the statistical prior distributions so that we can speed up convergence and reduce the risk of overfitting the model to particular individuals. Results from the ablation test validate its effectiveness.
- To validate the effectiveness of our method, the first multi-modal cognitive load and drowsiness driving dataset (MCDD) in the L3 context was created. In total, 42 subjects were involved in this dataset, which included about 105,840 seconds RGB video and cardiac and respiration-related physiological indicators. The cleaned processed dataset will be released.

II. Related Works

A. Physiological Signal-Based Driver State Monitoring

Driver drowsiness [41] can negatively affect drivers' performance when driving with automation, especially in takeover events. Thus, driver state monitoring is necessary to ensure safe driving in the context of driving automation [12]. According to [42], [43], in conditional autonomous vehicles, certain physiological metrics can be strongly correlated with driver state. Therefore, researchers proposed various physiological signal-based driver state monitoring methods. Commonly used psychophysiological measures include electrophysiological signals and human physical movements [44]. For example, Cui et al. [45] designed an interpretable convolutional neural network for drowsiness detection using EEG signals. In addition, due to the difficulty of obtaining EGG signals, methods based on Electrocardiography(ECG) signals [46] and Electrodermal Activity(EDA) signals [47] have also been widely proposed. Yang et al. [35] developed an attention-enabled recognition network with a decision-level fusion architecture to estimate cognitive load based on physiological data. Another study used physiological signals and facial images to detect driver drowsiness [21]. However, the above-mentioned studies used physiological data collected through intrusive sensors, i.e., the sensors had to contact with the driver's body, which makes them infeasible to be used in the actual driving environment. Therefore, the VDMoE proposed in this paper uses a non-contact RGB camera as the perception method and simultaneously presents multiple driver states to replace the traditional driver monitoring based on contact sensors.

B. Remote Physiological Measurement

Since 2008, Verkruysse et al. [36] proposed the rPPG technique for detecting physiological data using only a single consumer-grade camera. Nowadays, the rPPG technique has become a mainstream method in remote physiological data monitoring because of its non-contact characteristics. For example, to improve the robustness of pulse rate measurements, Hann et al. [48] proposed an analytical approach to tackle the motion problem in rPPG measurement. Wang et al. [49] introduced a mathematical model to increase understanding of the rPPG and designed a new algorithm to measure the heart rate. Tarassenko et al. [50] used autoregressive modeling to implement respiratory rate monitoring. However, traditional rPPG methods require manual adjustment of inaccurate data and are mostly limited to a single task.

Recently, the combination of deep learning algorithms and rPPG technology has effectively improved accuracy in complex environments [51], [52]. Many studies have been conducted in the direction of multi-task monitoring of physiological data. For example, Liu et al. [53] presented a multi-task network to enable respiration and heart rate measurements on a mobile platform. Narayanswamy [54] et al. proposed a model based on facial movement, heart rate and respiration measurement. Wang et al. [28] designed a PhysMLE model to simultaneously calculate HR, heart rate variability, respiration rate, and blood oxygen saturation. Nevertheless, existing multi-task rPPG approaches typically focus on physiological signal estimation (e.g., HR and RR) or single psychological indicators, rather than jointly inferring both physiological and cognitive/drowsiness states. Consequently, they lack the robustness and comprehensive capability required for real-time driver monitoring under Level-2/3 automation, which our VDMoE model explicitly addresses.

C. Non-contact Driver State Monitoring

Over the past few decades, with the development of deep learning technologies and increased demands for driving safety, non-contact driver state monitoring has become mainstream. Non-contact vision-based detection is categorized into conventional criterion-based methods and direct deep-learning detection methods. Specifically, using criteria combined with traditional methods such as eyes opening degree, mouth opening degree, or head posture, researchers were able to successfully detect driver states [55]. However, these criterion-based approaches are difficult to apply in diverse complex driving situations. Hence, direct drowsiness detection algorithms based on deep learning algorithms have become a more feasible solution. For example, Ma et al. [56] proposed an image-based yawn detection algorithm that uses the driver's image as an input for drowsiness detection. However, this method can only detect drowsiness by a single static image while losing temporal information in continuous video streams. Thus, video-based deep-learning models have also been proposed [57], [58], [59]. Although these methods are effective in obtaining time information, there are additional training and deployment costs. Recently, Mou et al. [20] developed a self-supervised learning algorithm using eye, nose, and head optical flow for drowsiness detection. In addition, some studies [55], [60] used facial landmarks to detect driver drowsiness. For example, Yang et al. [18] used eye, mouth, and facial feature points as inputs to determine the level of driver drowsiness.

While recent video-based DMS and rPPG research have begun exploring multi-task architectures, particularly those based

on deep CNNs or Transformer backbones, these approaches still face substantial parameter overhead as tasks are added. Existing multi-task rPPG methods often suffer from model bloat due to task-specific decoupling layers and conflicting optimization objectives, leading to a 'seesaw effect' between tasks [28], [54]. Moreover, to our knowledge, no non-contact video-based algorithm has concurrently addressed cognitive load assessment, drowsiness detection, and physiological monitoring in Level-2/3 driving. In contrast, VDMoE leverages a MoE-like design that is inherently well-suited for multi-task learning: a small set of shared expert MLPs is dynamically routed by an element-wise gating network, and spatio-temporal feature extraction is separated across specialized experts. This architecture allows new tasks to be accommodated without a proportional increase in overall parameters. By adopting lightweight two-layer MLP experts instead of heavy CNN or Transformer modules [52], [53], VDMoE achieves notably smaller model size and lower inference latency while also integrating rPPG to capture physiological information. Consequently, VDMoE not only outperforms existing Transformer- or CNN-based multi-task models in efficiency and accuracy, but also uniquely provides a fully non-contact solution for simultaneous estimation of driver drowsiness, cognitive load, and physiological states in Level-2/3 automation contexts.

TABLE II
A SUMMARY OF SYMBOLS AND DESCRIPTIONS

Symbol	Description
B	Number of samples in one batch.
T	Number of frames of each video.
N_f	Number of facial landmark points in each frame.
x_i, y_i	The horizontal and vertical coordinates of a facial landmark.
L, W, C	Length, width, and the number of channels of facial regions.
L_e, W_e	Length, width of facial regions of eyes.
L_m, W_m	Length, width of facial regions of mouths.
I_l, I_r, I_m	The local area of left eyes, right eyes, and mouths
N_s	Number of ROIs of STMaps.
I_l	The local facial regions of drivers.
I_s	STmaps.
I_f	The facial landmarks matrix.
X	Model inputs.
$f(*)$	The target multi-task model.
θ	Learnable parameters in the model.
$P(*)$	The task-specific estimation heads.
h_i	The ground-truth HR .
r_i	The ground-truth RR.
S	Size of the sliding step.
L	Number of VDMoE block.
K	Number of experts in the VDMoE block.
D	Spatial dimension of each feature embedding.
V	Spatial dimension of the combined feature embedding.
m_t	Feature vector maintains the temporal information.
m_s	Feature vector with spatial information.
$E(*)$	The expert network.
$R(*)$	The expert router network.
$G(*)$	The task-related feature gating network.
τ	Temperature factor for soft regularization.
λ	Adaptation factor for regularization term.
k	Trade-off parameter.

III. METHODOLOGY

A. Problem Formulation

From previous studies in human factors [26], [27], we know that both drivers' drowsiness and cognitive load are not only highly correlated with each other, but are also significantly associated with facial features (e.g., the degree of opening and closing of the eyes and mouth) and physiological features (e.g., HR, RR). Therefore, integrating state and physiological estimation into one multi-task model can effectively exploit correlations between tasks and features while compressing the computational cost further.

Inspired by this, we introduced a multi-task learning architecture, a video-based multi-task DMS mixture of experts model (VDMoE). As shown in Figure 3, suppose we have a batch of B raw videos. Being different from previous methods taking 3D videos as input [57], [20], we extracted heterogeneous facial features from videos as input X. Firstly, following [18], we extracted facial landmarks, and subregion facial videos (i.e., eyes and mouth) from the video. Then, as only movement features from landmarks and above subregions might be insufficient for physiological measurement, we combined the knowledge from rPPG [61], [62] to supplement skin light changes into the input X. The specific preprocessing steps are in the following section. Finally, given X, we tried to obtain both physiological signals $\{h_i, r_i\}_{i=1}^{N}$ and state estimation $\{d_i, c_i\}_{i=1}^{N}$ by training a multi-task model $f(X;\theta)$, where h_i is the heart rate, r_i is the respiratory rate, d_i is drowsiness levels and c_i is the cognitive load levels. The trainable parameter θ is our optimization objective. Finally, the multi-task estimation target is $Y = \{h, r, d, c\}$

B. Preprocess

In the original dataset, each video consists of T frames. Since each input sample is required to be of the same size, we segmented and augmented the raw dataset with the sliding window to obtain more samples with the same frames, effectively alleviating overfitting. We set the number of frames of all video samples to a fixed value T, and obtain the samples from the original video through a sliding window with a step size of S.

At the beginning of preprocessing, we first applied the facial detection methods to each frame of the input video to generate 2D coordinates of N_f facial landmarks with N_f key points. Since the length of each video is T frames and each facial landmark is represented by horizontal and vertical coordinates, denoted as (x_i, y_i), each video can be represented as a facial landmarks matrix of shape $T \times N_f \times 2$. Then, by using the outer boundaries corresponding to the facial landmarks as bounding boxes, the regions of the left eye, right eye, and mouth I_l, I_r, I_m were obtained from each original facial video frame. To standardize the input, the subregion videos of the left eye and right eye I_l, I_r are resized to $T \times L_e \times W_e \times C$, where L_e, W_e are the length, width of the subregion videos corresponding to eyes. Similarly, the subregion video of mouth I_m is the size of $T \times L_m \times W_m \times C$, where L_m, W_m are the length, and width of the mouth video. Subsequently, the entire facial areas were cropped into N_s regions of interest (ROIs) [63]. These ROIs were converted from RGB to YUV. Then, since we know that the skin light changes owing to the volume and oxygen saturation of the blood correspond to the low amplitude of the low-frequency

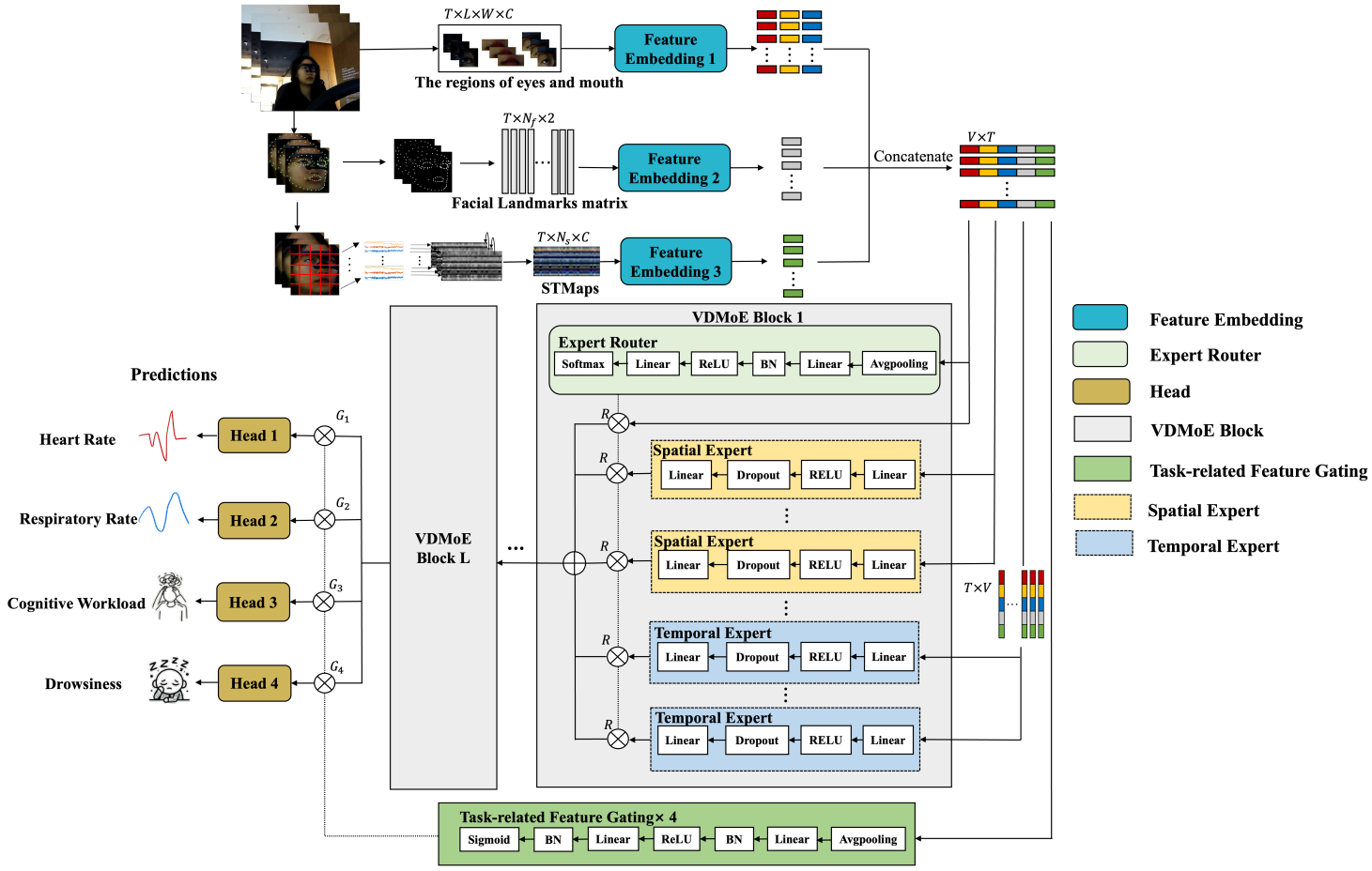

Fig. 1. This figure presents the proposed multi-task driver state monitoring method VDMoE. We first transform videos into STMaps, facial landmarks matrix, and facial regions. Then inputs them into the model to output physiological signals, drowsiness, and cognitive load.

component [64], we used the first-order difference to remove the large amplitude caused by head movements and changes in light. After a Butterworth filter with parameter [0.4,10] Hz to retain the frequency band of interest [63], we processed the ROIs into a spatial-temporal map (STMap) $I_s \in \mathbb{R}^{T\times N_s\times C}$ to measure physiological signals. Among them, the channel number $C = 3$ is constant. In all, the model's input X can be described as $X = \{I_l, I_r, I_m, I_f, I_s\}$.

C. VDMoE

1) Feature Embedding Modules: Due to the different shapes of the heterogeneous facial features (i.e., STMap I_s, facial landmarks matrix I_s, and the regions of the left eye I_l, right eye I_r, and mouth I_m), we designed feature extraction modules tailored to ensure that facial features from different types of inputs are encoded to the same size of high dimension space. The specific architecture and parameters of the network are listed in Table III. Specifically, a series of convolution and pooling operations for each frame of the facial subregion videos were built separately, simultaneously processing feature representations from different temporal frames and maintaining the temporal information. First, we performed a dimension-up operation and then a convolution to keep the temporal order of the matrix. Then, for the feature embedding of STMaps, due to the characteristics of STMaps that condense both temporal and spatial information, we designed a series of convolutions, which retained the dimensions of STMaps in the spatial dimension, and kept the correlation between the temporal features of STMaps and other input features. At the same time, in the final feature embedding stage, we adopted adaptive pooling and reshaping steps to compress the spatial dimension information into one-dimensional vectors, aligning the output dimensions with those of other feature embeddings. The facial landmarks matrix was formed by combining the facial landmarks in chronological order per frame. In other words, we used a feature embedding similar to the facial region for each frame of facial landmarks. Finally, we compressed the features into one-dimensional vectors through a fully connected layer to obtain a fixed output shape.

Through feature embedding, each type of input facial feature was converted to a two-dimensional vector of shape $T \times D$. It is worth noting that, since all dimension transformations were conducted over the spatial dimension (i.e., dimensions exclude the T for each sample) and we fixed the temporal dimension T in the embedding phase, our feature embedding process actually distilled the spatial facial features of each frame into a vector of size D and preserved the temporal ordering. Finally, by stacking these vectors at the last dimension, we got the combined feature vector m with the shape of $T \times 5D$. For descriptive convenience, we generalize $5D$ to spatial dimension V.

2) VDMoE Block: To learn the underlying correlations between different tasks, based on the classic MoE block [37], we proposed the hierarchical VDMoE architecture to accommodate the spatio-temporal characteristics of heterogenous facial features. There are L VDMoE blocks in total. Specifically, as previous feature embedding modules have already encoded heterogeneous facial features to high-dimensional space, to minimize the computational cost and avoid the decline in generalizability due to high complexity, we instantiated each expert in the VDMoE block with a two-layered MLP. Additionally, given our m put the spatial information from heterogenous facial features at the last dimension, following the matrix multiplication in linear algebra, the weights of MLP experts can only be applied to the last dimension (i.e., spatial information). However, the temporal information is also important [55], as both drivers' behavior and physiological reflection under specific states should not be instantaneous. Thus, we introduced the spatio-temporal expert mechanism in our VDMoE block. Specifically, given that only spatial information can be processed in m, we first transposed the m at the last two dimensions and obtained the $m_t \in \mathbb{R}^{V\times T}$, where temporal information of each spatial channel can be learned. Then, in each VDMoE block, we constructed K experts, which can be further equally classified as spatial experts $E_s(*)$ and temporal experts $E_t(*)$ corresponding to m and m_t respectively. Experts share no parameters and are initialized separately. The structure of each type of expert is shown in Figure 1. Then, the m was processed to $\frac{K}{2}$ $m_s^{'}$ by $\frac{K}{2}$ different spatial experts $E_s(*)$. Similarly, we obtained $\frac{K}{2}$ $m_t^{'}$ after input m_t to $\frac{K}{2} E_t(*)$. Moreover, to avoid information degradation with the increase in the number of layers in neural networks [39], we leveraged the jump connection in the residual network. Therefore, the intermediate features $m^{'}$ consist of $\frac{K}{2}$ $m_s^{'}$, $\frac{K}{2}$ $m_t^{'}$, and the initial input of the VDMoE block m. The whole procedure is formalized as follows:

$$\begin{aligned} &m \in \mathbb{R}^{T\times V}, \quad m_t = m^{\top} \in \mathbb{R}^{V\times T}, \\ &m_s^{'} = \{E_s^i(m)\}^{\frac{K}{2}}, \ i \in [1, \frac{K}{2}], \\ &m_t^{'} = \{E_t^i(m_t)\}^{\frac{K}{2}}, \ i \in [\frac{K}{2}, K], \\ &m^{'} = \{m_s^{'}, m_t^{'\top}, m\}. \end{aligned} \tag{1}$$

Then, we designed an expert router mechanism $R(*)$ to dynamically aggregate the above intermediate features. Each router $R(*)$ corresponds to a specific intermediate feature. The router $R(*)$ took the m as the input and nonlinearly converted it from $\mathbb{R}^{T\times V}$ to a probability vector of shape $(K+1)\times 1$. In detail, upon receiving the m, we first applied average pooling to compress the temporal dimension. Then, a two-layer MLP performed a nonlinear transformation, followed by a Softmax function to generate the output. Finally, the probability vector generated by the router was used to selectively activate features by weighting the intermediate variables. Each element of the vector represents the activation probability for a corresponding feature. By weighted summing these intermediate features with probability features, we emphasized important features while suppressing less relevant ones. The aggregated feature $m^{'}$ was sent to the next VDMoE block.

3) Task-Related Feature Gating and Estimation Head: After the VDMoE blocks, a representative task-generic feature $m^{'} \in \mathbb{R}^{T\times V}$ was obtained. Although there are correlations among multiple target tasks and can be learned by one shared backbone, separate feature spaces for different tasks are still necessary [28]. Therefore, we designed element-wise gated networks to select sparse high-level representations for specific tasks. Specifically, four element-wise gates $G_i(*)$ were initiated, which took the original embedding feature m as input. Then, seeing Figure 1, after the average pooling and a two-layered MLP, the output of each gate $G_i(*)$ was controlled by a Sigmoid function, which makes each element of the output vector in the range [0, 1]. We argue that, being different from Softmax-based weighted aggregation [37], [65], the element-wise Sigmoid-based gate can adaptively control the importance of low-level information relative to specific tasks, thus selecting proper subspaces $\{m_i^{'}\}^4 \in \mathbb{R}^V$ from a uniform task-generic space $m^{'}$. Lastly, we input these subspaces into four task-specific estimation heads as:

$$\begin{aligned} Y &= P(G(m) \cdot Pool(m^{'})) \\ &= Linear(Sigmoid(MLP(Pool(m))) \cdot Pool(m^{'})). \end{aligned} \tag{2}$$

D. Prior-driven Alignment

As noted by [26], drowsiness and cognitive load tend to exhibit an overall negative correlation in driving contexts. This inverse relationship is rooted in cognitive-resources theory [66], which indicates the limited human attentional capacity: when task demands are low, surplus resources 'wander', leading to mind-wandering and increased drowsiness [67]. Conversely, moderate task demands optimally engage attentional control mechanisms, sustaining cortical arousal, maintaining vigilance, and delaying the onset of fatigue [68]. Studies of vigilance decrement further demonstrate that understimulation produces lapses in attention and microsleeps [69], whereas appropriately challenging tasks improve sustained performance. We therefore leverage this well-established theoretical and empirical foundation to introduce a prior-driven soft regularization $\mathcal{L}_{align}$, enforcing a general negative alignment between our drowsiness and cognitive workload estimates

Specifically, given the estimated vector of drowsiness and cognitive load by corresponding heads, we applied the Softmax function to each vector to convert them to the binary probability vector p_{drow} and p_{cog}. Then, based on the Kullback-Leibler Divergence, we tried to maximize the distributional difference between p_{drow} and p_{cog} with Eq. (3). Additionally, considering this correlation might not be robust in some extreme situations (e.g., cognitive load exceeding the cognitive capacity will also speed up drowsiness development [70]), we applied a temperature hyper-parameter τ to scale the two probability distributions, which can control the sharpness of the output distribution. After trial and errors, we set τ as 0.5 to smooth the distributions:

$$\mathcal{L}_{align} = -\sum_{B} \exp\left(\log\left(p_{drow}/\tau\right)\right) \log\left(\frac{\exp\left(\log\left(p_{drow}/\tau\right)\right)}{p_{cog}/\tau}\right) \quad (3)$$

TABLE III
NETWORK STRUCTURES OF FEATUREEMBEDDING OF STMAPS, FACIAL LANDMARKS MATRIX AND FACIAL REGIONS

Layer	Input size	Output size	Content
Feature Embedding 1 of Facial Regions			
Conv1	$B \times H \times W \times C$	$B \times H \times W \times 16$	Conv + ReLU
Pool1	$B \times H \times W \times 16$	$B \times \frac{H}{2} \times \frac{W}{2} \times 16$	MaxPool
Conv2	$B \times \frac{H}{2} \times \frac{W}{2} \times 16$	$B \times \frac{H}{2} \times \frac{W}{2} \times 32$	Conv + ReLU
Pool2	$B \times \frac{H}{2} \times \frac{W}{2} \times 32$	$B \times \frac{H}{4} \times \frac{W}{4} \times 32$	MaxPool
Reshape	$B \times \frac{H}{4} \times \frac{W}{4} \times 32$	$B \times (\frac{H}{4} \times \frac{W}{4} \times 32)$	Reshape
FC	$B \times (\frac{H}{4} \times \frac{W}{4} \times 32)$	$B \times 128$	Fully Connected
Feature Embedding 2 of STMaps			
Conv1	$B \times N_s \times C$	$B \times N_s \times 64$	Conv + ReLU
BatchNorm1	$B \times N_s \times 64$	$B \times N_s \times 64$	BatchNorm
Conv2	$B \times N_s \times 64$	$B \times N_s \times 128$	Conv + ReLU
AdaptivePool	$B \times N_s \times 128$	$B \times 128$	AdaptivePool
Reshape	$B \times 128$	$B \times 128$	Reshape
Feature Embedding 3 of Facial Landmarks Matrix			
Reshape1	$B \times N_f \times 2$	$B \times 1 \times N_f \times 2$	Reshape
Conv1	$B \times 1 \times N_f \times 2$	$B \times 32 \times N_f \times 1$	Conv + ReLU
BatchNorm1	$B \times 32 \times N_f \times 1$	$B \times 32 \times N_f \times 1$	BatchNorm
Reshape2	$B \times 32 \times N_f \times 1$	$B \times (32 \times N_f)$	Reshape
FC1	$B \times (32 \times N_f)$	$B \times 128$	Fully Connected

E. Optimization Goal

To facilitate the estimation of drowsiness, cognitive load, HR, and RR. We designed different optimization goals depending on the properties of each task. For the loss of drowsiness $\mathcal{L}_{drow}$ and cognitive load $\mathcal{L}_{cog}$, given there is strong subjective consciousness and individual differences in assessing drowsiness and cognitive load with the self-report questionnaires [71], participants will label their current status with varied standards. When training the model with data from different individuals, the mapping between facial features and state labels is not robust. Thus, we applied the generalizable cross-entropy loss (Truncated Loss) [72] to optimize the drowsiness and cognitive load estimation, and the smooth L1 loss for the estimation of HR $\mathcal{L}_{hr}$ and RR $\mathcal{L}_{rr}$ following [61]. Besides, to suppress some meaningless effects of the regularization at the early iterations, we applied the adaptation factors λ as Eq. (4)

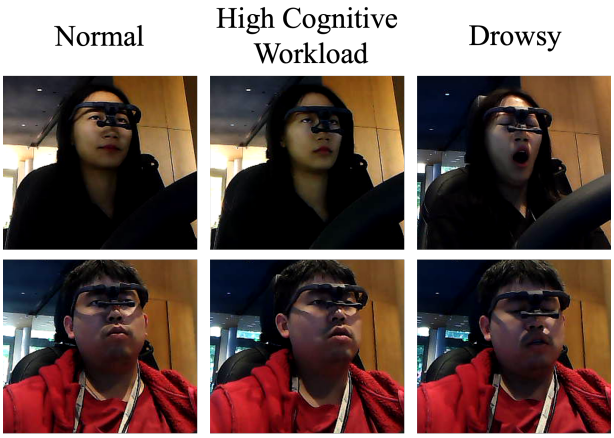

Fig. 2. Sample RGB video frames of drivers in different states in MCDD.

$$t = \frac{Iter_{current}}{Iter_{total}}, \quad \lambda = \frac{2}{1 + \exp -10t}. \quad (4)$$

Then, we combined the multi-task optimization goals and the regularization term into one loss formula $\mathcal{L}_{overall}$ to conduct the joint training with one trade-off parameter k_1.

$$\mathcal{L}_{overall} = \mathcal{L}_{drow} + \mathcal{L}_{cog} + \mathcal{L}_{hr} + \mathcal{L}_{rr} + \lambda * k_1 * \mathcal{L}_{align}. \quad (5)$$

IV. MATERIALS

A. MCDD Dataset

To mitigate the gap in previous public datasets, we collected a new multi-modal cognitive load and drowsiness driving dataset (MCDD). Some example frames of drivers with different states and the distribution of our extracted HR and RR are shown in Figure 2.

1) Experiment design: Previous study [73] showed that drivers' cognitive resources are multi-dimensional, and different dimensions of non-driving-related tasks (NDRTs) can result in different levels of cognitive load. Given the moderating effect of drivers' cognitive load on drowsiness and physiological responses [8], [27], this driving simulator study used a within-subject design to investigate the impact of varying types of cognitive NDRT on drivers' states. As detailed in Table V and Figure 3, three types of cognitive tasks, encompassing 6 specific tasks, were used in addition to a baseline condition without NDRTs, leading to 7 NDRT tasks. Each participant completed three trials per NDRT task, resulting in a total of 21 drives (7 NDRTs * 3 trials). A Latin square design, with 21 unique orders, was implemented to mitigate potential order effects.

All driving scenarios took place on simulated two-way, six-lane highways with a speed limit of 120 kilometers per hour and a traffic density of 6 vehicles per kilometer per lane. The driving automation system was set to keep a speed of 110 kilometers per hour. During manual driving segments, participants were instructed to remain in the middle lane. Each drive covered a distance of approximately 7 kilometers.

TABLE IV
SUMMARY OF COGNITIVE LOAD TASKS.

Task Type	Description	Example	Cognitive Load Level	Cognitive Resource
N-back Task [74]	A series of stimuli numbers are presented with a pause between each. Participants recall and verbally report the stimulus that is n positions earlier.	0-back (0-B), 1-back (1-B), 2-back (2-B) tasks.	Three levels; The difficulty increases with the increase of n.	Memory
Math Task [11]	Oral backward counting from 3,000 by increments of 3 (non-integer) or 5 (integer).	Counting backward by 3 (MT1) or 5 (MT2) from 3,000.	Two levels; level of cognitive depends on the nature of the math tasks.	Calculation
Spatial Task [75]	Participants listen to an audio clip describing a route and identify the main direction faced at the end.	E.g., "What direction is this person when he goes to the north station and moves two stations clockwise?" (Answer: East) (ST)	One level: Simulates high cognitive demands similar to those in navigation systems.	Spatial Processing

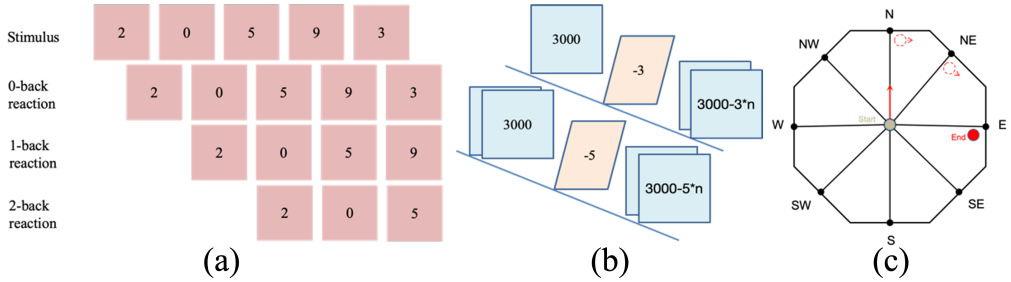

Fig. 3. Cognitive task design, including (a) n-back task, (b)math task, and (c) spatial task.

Participants were provided with pre-experiment instructions to ensure adequate rest and avoid substances that could impact their performance. These instructions, given 24 hours prior to the experiment, included maintaining regular sleep patterns, abstaining from alcohol, and refraining from caffeine intake. Upon arrival, participants provided written informed consent and then participated in a 30-minute training session. This session covered the experimental procedure, vehicle operation, cognitive tasks, and subjective questionnaires. Before commencing the experimental drives, physiological sensors were fitted and calibrated. After each drive, participants completed questionnaires to assess their cognitive load using the NASA Task Load Index (NASA-TLX) [76] (ranging from 1 to 20) and drowsiness level using the Karolinska Sleepiness Scale (KSS) [77] (ranging from 1 to 10). The total experiment duration for each participant was approximately 3 hours, with about 1.5 hours dedicated to driving. This study was approved by the Human and Artefacts Research Ethics Committee at the Hong Kong University of Science and Technology (protocol number: HREP-2023-0199).

2) Participants: Before conducting the experiment, a power analysis was performed using MorePower software [78] to determine the minimum sample size needed. The analysis indicated that a sample size of 24 participants would be sufficient to generate a statistical power of 80%, with confidence interval = 95% and effect size of (σ^2) of 0.06.

The study involved a total of 42 participants (25 male, 17 female) with a mean age of 35.28 years (SD = 9.10, range = 23-53 years). Participants were recruited across four age groups (20-60 years) to minimize the influence of age-related factors on physiological signals and enhance the model's generalizability. All participants were required to have a valid driver's license for at least one year and no prior experience with advanced driving systems (ADS). Participants received

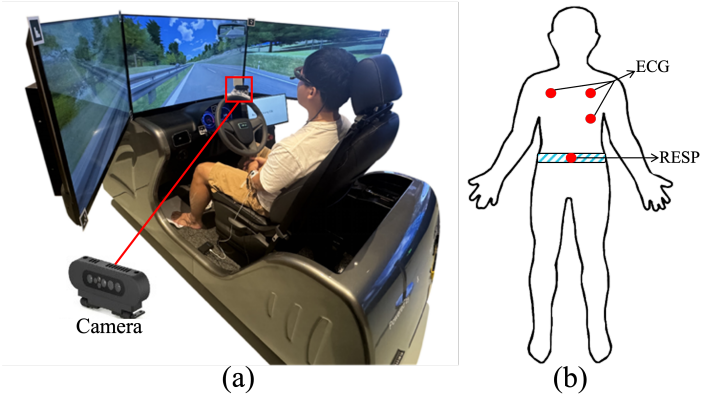

Fig. 4. (a) The driving simulator; (b) Data collection sensors.

an hourly compensation of 70 RMB and were given the opportunity to earn a performance-based bonus of up to 30 RMB for completing cognitive tasks.

3) Apparatus: The study utilized a fixed-base driving simulator equipped with three 42-inch screens, providing a 150° horizontal and 47° vertical field of view (Figure 4(a)). An external tablet with two touch buttons allowed participants to activate and deactivate the driving automation system. Driving scenarios were developed and vehicle operation data was collected at a frequency of 60 Hz using Silab 7.1 software by WIVW.

Participants' facial RGB videos were recorded using an Orbbec Gemini pro camera. The camera was installed in front of the driver, which was recorded at a resolution of 640x480 pixels and a frequency of 30 Hz. Physiological data, including electrocardiogram (ECG), and respiration (RESP), were collected at a frequency of 100 Hz using a 3-channel SMD electrocardiograph and respiratory belt from Ergoneers (Figure 4(b)), based on which the ground truth HR and RR

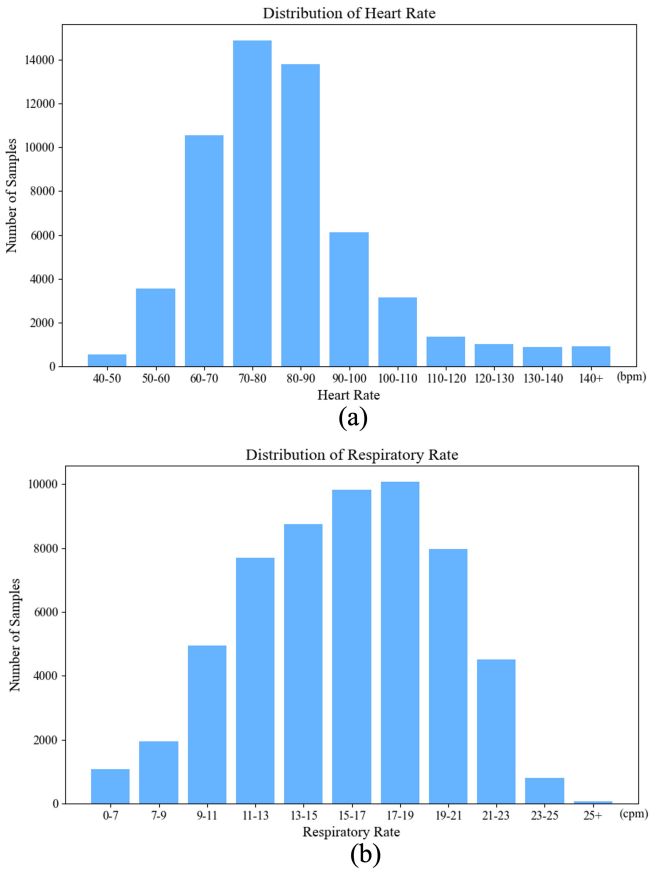

Fig. 5. The distribution of HR and RR in MCDD.

were calculated. The distribution of HR and RR in the MCDD dataset is visualized in Figure 5.

Besides, following [79], based on the KSS questionnaire collected after each trial, there are 75.86% samples labeled as 'awake' (KSS <5), and 24.14% labeled as 'drowsy' (KSS $\geq$ 8). For cognitive workload, we first normalized each participant's raw NASA-TLX scores to the 1–20 range to mitigate inter-subject differences [80]. We then dichotomized the normalized scores at a global midpoint (10), yielding 29.61% of trials labeled as 'high cognitive load' (score >10) and 60.39% as 'normal cognitive load' (score ≤ 10). A sensitivity analysis shifting the threshold by ±2 points showed model performance variance below 5%, confirming that our median-split strategy produces robust class balances without unduly biasing results.

B. Other Datasets

We also evaluated our method in two other public datasets. However, as there is no dataset that satisfies the requirements of the multi-task remote detection (i.e., more than one driver state, and contains physiological signals), we selected one dataset for detecting drowsiness (FatigueView) [33] and one another dataset for multi-task vital signs (V4V) [81].

Specifically, the FatigueView dataset consists of videos captured by RGB and infrared (IR) cameras from five different locations. These videos depict real-life scenarios of drowsy driving and tagged visual cues of drowsiness, varying from subtle to obvious. As no physiological signals or cognitive load levels were collected in this dataset, we utilized the RGB-Front video only from this dataset. In addition, the V4V dataset was widely used in rPPG [61], [28]. It collected physiological signals and facial videos under ten tasks and the ground-truth HR and RR were provided in V4V; however, it was not driving-related and no driver states were assessed.

V. Experiment

A. Implementation Details

For the experiments on three datasets, the cross-subject evaluation protocol was utilized. Specifically, for all subjects in each dataset, we randomly separated them into training, validation, and test sets with the ratio 6:2:2. Models were trained on the data from subjects in the training set, adjusting parameters in the validation set, and only reporting the results in the test set. Before training, to ensure a more accurate RR and HR measurement, we set the sliding window size $F = 300$ frames and step size $S = 30$ frames. Then, the MediaPipe Face Mesh package from Google[1] was leveraged to perform landmark detection. Following [18], we selected 106 facial landmarks. The size of landmark matrix I_f is $\mathbb{R}^{300\times106\times2}$. To standardize the input, each frame of the left eye and right eye I_l, I_r was resized to $\mathbb{R}^{25\times25\times3}$ and the mouth frame was resized to $\mathbb{R}^{35\times15\times3}$. Additionally, based on 106 landmarks of each sliding window, we further generated STMaps following [85]. The size of STMap is $\mathbb{R}^{300\times25\times3}$. Then, in Table III, we present the specific configuration of our feature embedding modules. In practice, we set up $K = 4$ experts in each VDMoE block, and there are $L = 1$ layers of VDMoE. Adam optimizer with a learning rate of 0.00001 was used for training. The trade-off parameter k was 0.001. The batch size B and maximum iterations were set to 250 and 20000, respectively.

Following existing works [86], [21], we used Accuracy, F1 score, Sensitivity, and Specificity to evaluate performance on cognitive load and drowsiness estimation. For HR and RR estimation, mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (p) were used [87]. We conducted five evaluations using five different random seeds, and then used a paired t-test to determine the significance of the mean performance difference between the best and second-best models. The whole work was implemented in the Pytorch framework and all experiments were conducted on an Nvidia RTX A6000.

B. Baseline Models

To evaluate the performance of our proposed VDMoE, we compared it with several single-task or multi-task models for both state and physiological estimation. Specifically, for single-task models, DBN-HMM [30], 3DCNN+BiLSTM [57], DDDNet [58], IsoSSL-MoCo [20], ReNeXt3D-101+LSTM [59], FacialUnits [60], 2s-STGCN [55], MIGCN [82], and VBFLFF [18] were used for drowsiness estimation, and CHROM [48], POS [49], ARM-RR [50], Dual-GAN [51], ConDiff-rPPG [62], and PhysFormer++ [52] were used for physiological measurement. For multi-task models, since there

[1]https://github.com/google-ai-edge/mediapipe

TABLE V
PERFORMANCE OF DROWSINESS AND COGNITIVE LOAD ESTIMATION ON MCDD.

Method	Drowsiness				Cognitive load			
	Accuracy↑	F1 Score↑	Sensitivity↑	Specificity↑	Accuracy↑	F1 Score↑	Sensitivity↑	Specificity↑
DBN-HMM$^+$ [30]	71.96	51.57	57.25	78.63	—	—	—	—
3DCNN+BiLSTM$^+$ [57]	72.18	51.09	52.84	79.86	—	—	—	—
DDDNet$^+$ [58]	74.22	51.09	51.31	82.10	—	—	—	—
IsoSSL-MoCo$^+$ [20]	76.82	56.32	59.62	87.71	—	—	—	—
ReNeXt3D-101+LSTM$^+$ [59]	78.61	63.30	61.32	87.45	—	—	—	—
MIGCN$^+$ [82]	70.24	50.75	59.31	83.09	—	—	—	—
FacialUnits [60]	73.78	52.21	53.07	79.35	—	—	—	—
2s-STGCN [55]	79.43	56.60	44.05	94.92	—	—	—	—
VBFLFFA [18]	81.05	65.39	58.81	90.78	—	—	—	—
ResNet3D$^+$ [83]	80.30	65.82	62.28	88.19	72.04	61.10	68.33	73.80
ViViT$^+$ [84]	79.81	60.62	51.05	92.40	71.76	56.57	57.25	78.63
VDMoE	**84.31***	**69.84***	**66.72***	**92.58***	**79.96***	**68.81***	**77.13***	**80.77***

Notes: In this and following tables, the mean of 5-time evaluations is presented. '—' means there are no evaluation results as the model cannot conduct the corresponding task. The **bold** shows the best result within each column, and the $*$ indicates the significantly (p-value $<$ 0.05 with the paired-t test) best result within each column. $+$ indicates methods taking facial video as models' input.

TABLE VI
PERFORMANCE OF HR AND RR ESTIMATION ON MCDD.

Method	HR			RR		
	MAE↓	RMSE↓	p↑	MAE↓	RMSE↓	p↑
CHROM$^+$ [48]	18.33	19.70	0.20	—	—	—
POS$^+$ [49]	17.33	20.02	0.22	—	—	—
ARM-RR$^+$ [50]	—	—	—	7.32	9.16	0.11
Dual-GAN [51]	13.29	18.86	0.31	—	—	—
ConDiff-rPPG [62]	15.32	19.58	0.29	—	—	—
PhysFormer++$^+$ [52]	13.15	18.33	0.34	—	—	—
MTTS-CAN$^+$ [53]	13.96	18.31	0.30	5.22	6.68	0.37
BigSmall$^+$ [54]	13.13	18.02	0.32	5.26	6.72	0.37
PhysMLE [28]	12.03	17.11	0.46	5.12	7.03	0.38
ResNet	13.46	20.24	0.31	5.44	7.07	0.37
ViT	13.00	19.69	0.35	5.04	6.94	0.39
ResNet3D$^+$ [83]	14.27	19.02	0.29	5.63	6.58	0.14
ViViT$^+$ [84]	14.92	19.61	0.28	5.33	6.10	0.16
VDMoE	**10.32***	**15.37***	**0.53***	**4.98***	**6.53***	**0.45***

is currently no facial video-based method proposed for multi-task state and physiological estimation, except for three methods for multi-task physiological estimation (i.e., MTTS-CAN [53], BigSmall [54], and PhysMLE [28]), we built four baselines with the hard-parameter sharing paradigm. In general, each multi-task baseline consists of a typical backbone network (i.e., ResNet3D and ViViT for 3D video input, ResNet and ViT for physiological estimation with STMap input [61]) and estimation heads that are the same as VDMoE.

Among the baselines, following their source papers, DBN-HMM, 3DCNN+BiLSTM, DDDNet, IsoSSL-MoCo, ReNeXt3D-101+LSTM, MIGCN, CHROM, POS, ARM-RR, PhysFormer++, MTTS-CAN, BigSmall, ResNet3D, and ViViT utilize only raw 3D facial video as input. In contrast, FacialUnits and 2s-STGCN rely on facial landmarks extracted from the facial video. Dual-GAN, ConDiff-rPPG, and PhysMLE use STMap as input, while VBFLFF incorporates landmarks along with videos of the eye and mouth subregions. The results are presented in Table V, VI.

C. Results of Comparison Experiment

Firstly, we identified that our proposed VDMoE significantly outperformed all baselines in both state and physiological estimation. Specifically, for single-task state estimation, we found that, methods taking facial video as input generally performed worse than those with preprocessed facial features in drowsiness estimation (e.g., the accuracy of ReNeXt3D-101+LSTM is worse than VBFLFFA by 7.2%). In addition, two multi-task baselines (i.e., ResNet3D and ViViT) outperformed other single-task methods that used facial videos (e.g., the accuracy of ResNet3D is higher than 3DCNN+BiLSTM by 11.2%). Nevertheless, our proposed VDMoE still achieved significantly better performance in both drowsiness and cognitive load estimation tasks than the best single-task model (VBFLFFA) by 4% and the second-best multi-task baseline (ResNet3D) by 5.1% in accuracy.

Secondly, referring to Table VI, VDMoE achieved the best performance on two physiological estimation tasks (HR and RR). In addition, compared to traditional single-task methods (i.e., CHROM, POS, ARM-RR), the rest deep methods have lower estimation errors in both two tasks. Moreover, for four multi-task baselines (i.e., ResNet, ViT, ResNet3D, and ViViT), baselines with facial videos performed worse in HR estimation than those taking preprocessed STMap. For instance, the MAE of ResNet is 5.7% lower than ResNet3D in HR estimation. Further, in RR estimation task, although baselines with facial videos sometimes had lower estimation error (e.g., the MAE of ViViT is lower than ResNet by 2%), their p was notably lower than most other deep methods and baselines (e.g., p of ResNet3D is lower than ViT by 64.1%).

D. Results of Ablation and Computational Cost Study

In this part, to elaborate on the effectiveness of VDMoE, we constructed several variants to compare with. In brief, except for four multi-task baselines (i.e., ViT, ResNet, ResNet3D, and ViViT), we constructed six variants by eliminating one facial feature and its corresponding feature embedding module,

TABLE VII
ABLATION STUDY ON DROWSINESS, COGNITIVE LOAD, HR, AND RR ESTIMATION WITH COMPUTATIONAL COST.

Method	Drowsiness		Cognitive load		HR		RR		Computational Cost	
	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑	MAE↓	p↑	MAE↓	p↑	Params(M)	FLOPs(G)
ViT	—	—	—	—	13.00	0.35	5.34	0.27	81.82	3.41
ResNet	—	—	—	—	13.46	0.31	5.44	0.27	12.74	2.21
ResNet3D$^+$ [83]	80.30	65.82	72.04	61.10	14.27	0.29	5.63	0.14	31.82	2.59
ViViT$^+$ [84]	79.81	60.62	71.76	56.57	14.92	0.20	5.33	0.16	9.92	5.56
VDMoE w/o I_l	81.65	61.44	74.47	58.41	11.72	0.42	5.32	0.38	6.08	1.47
VDMoE w/o I_r	77.08	56.34	73.11	61.95	11.90	0.45	5.15	0.43	6.08	1.47
VDMoE w/o I_m	79.49	60.58	72.42	59.01	11.43	0.43	5.25	0.39	6.12	1.53
VDMoE w/o I_f	80.69	64.17	74.20	62.91	11.17	0.41	5.30	0.30	6.15	1.24
VDMoE w/o I_s	81.05	64.26	74.69	63.39	12.72	0.40	5.13	0.41	5.81	1.64
VDMoE w/o $\mathcal{L}_{align}$	82.35	67.30	71.22	59.12	10.94	0.46	4.98	0.44	4.17	1.80
VDMoE	**84.31***	**69.84***	**79.96***	**68.81***	**10.32***	**0.50***	**4.98**	**0.45**	4.17	1.80

or the proposed prior-driven alignment regularization $\mathcal{L}_{align}$ from the complete VDMoE. The results are shown in Table VII. In addition to multi-task estimation performance, the computational costs (i.e., Parameter number and FLOPS) are also presented.

As shown in Table VII, the complete version of VDMoE significantly outperformed all variants without notable computational cost increases. Compared to baselines, VDMoE not only performed better in all tasks, but also with less computational cost (e.g., 23.3% increase in the MAE of HR estimation but with 69% parameter size of ResNet).

At the same time, we noticed a notable decrease in the performance of drowsiness estimation (9.4% in accuracy) when the I_r module was removed from VDMoE, and a decrease in cognitive load estimation when $\mathcal{L}_{align}$ was removed (12% in accuracy). In addition, for physiological estimation, the variant without I_s has the worst performance in HR estimation, while the lowest estimation precision of RR estimation belongs to the variant without I_f. Nevertheless, compared to baselines with facial video, most variants can achieve better multi-task estimation performance, particularly for physiological indicators, and maintain a relatively lower computational cost at the same time. For example, as the variant that performed poorest in HR estimation, the MAE of VDMoE without I_s was still lower than ResNet3D by 10.9%, even with only around 1/5 parameters. Compared to baselines that can only make physiological estimations (ResNet and ViT), all variants still dominate in physiological estimation task performance and computational costs.

E. Impact of K and L

The K denotes the number of experts in each layer of the model, while the L denotes the number of layers in the VDMoE. Since the K and L control the width and depth of the VDMoE, we conducted a hyper-parameter test to verify the best choice of K and L, and their impact on the multi-task performance. We visualized the results in Figure 6.

Firstly, upon examining the drowsiness and cognitive load estimation (Figure 6(a)(b)), there did not appear to be a consistent trend with respect to the number of K and L; however, it is notable that the highest F1 score for drowsiness and cognitive load detection was both achieved when $K = 4$, $L = 1$. This suggests that a moderate increase in the number of experts with a single layer is beneficial for these tasks. The Pearson correlation p coefficients in the HR and RR estimation (Figure 6(c)(d)) vary across different configurations, with some of the highest values observed with $K = 2$, $L = 2$ for HR and $K = 4$, $L = 1$ for RR. It is worth noting that configurations with $K = 2$ tend to lead to a higher association for HR, while $K = 4$ tends to perform better for RR.

TABLE VIII
PERFORMANCE OF HR AND RR ESTIMATION ON V4V.

Method	HR			RR		
	MAE↓	RMSE↓	p↑	MAE↓	RMSE↓	p↑
CHROM$^+$ [48]	11.44	16.97	0.28	—	—	—
POS$^+$ [49]	14.59	21.26	0.19	—	—	—
ARM-RR$^+$ [50]	—	—	—	8.11	12.23	0.08
Dual-GAN [51]	4.93	7.68	0.81	—	—	—
ConDiff-rPPG [62]	5.19	7.88	0.80	—	—	—
PhysFormer++$^+$ [52]	4.88	7.62	0.80	—	—	—
MTTS-CAN$^+$ [53]	5.31	8.03	0.79	1.98	6.31	0.15
BigSmall$^+$ [54]	5.03	7.84	0.81	1.88	5.83	0.17
PhysMLE [28]	4.79	8.06	0.79	1.55	5.40	0.21
ResNet	4.64	7.98	0.80	2.59	6.77	0.12
ViT	4.91	8.57	0.76	1.94	5.98	0.17
ResNet3D$^+$ [83]	8.32	11.70	0.61	2.03	6.09	0.09
ViViT$^+$ [84]	7.91	9.88	0.69	1.81	5.89	0.16
VDMoE	**4.01***	**6.82***	**0.85***	**1.02***	**4.94***	**0.31***

Notes: In this table, since there are no cognitive load and drowsiness estimation tasks, the VDMoE only applied the two losses for HR and RR regularization.

Additionally, there is a general trend that increasing the number of experts and layers leads to a larger number of parameters, which is consistent with the expectation that more complex models have a higher capacity (and potentially greater computational demands). In general, there is no clear linear correlation between the number of parameters and the performance metrics (F1 scores and Pearson correlations). For instance, the highest number of parameters (15.39 M, see 6(e)) does not correspond to the highest performance across tasks, indicating that simply increasing model complexity does not guarantee improved performance.

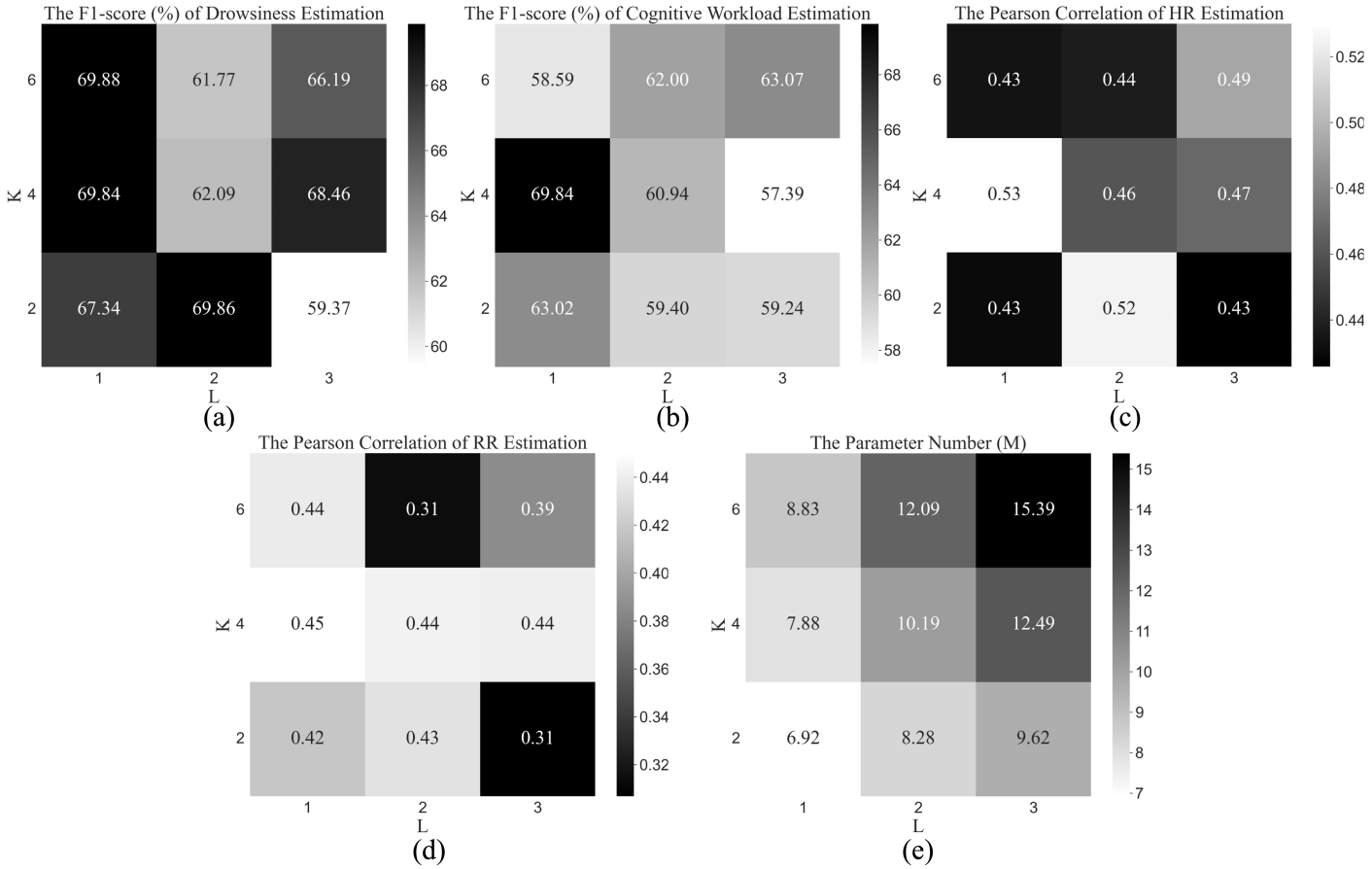

Fig. 6. The visualization of the impact of hyper-parameter K and L to each task. Subfigure (a)(b) indicates the F1-score of drowsiness and cognitive load estimation, (b)(d) is the Pearson correlation of HR and RR estimation, and (e) presents the parameter number (M) of VDMoEs with different K and L.

TABLE IX
PERFORMANCE OF DROWSINESS ESTIMATION ON FATIGUEVIEW.

Method	Accuracy↑	F1-score↑	Sensitivity↑	Specificity↑
DBN-HMM$^+$ [30]	79.35	81.66	86.35	80.61
3DCNN+BiLSTM$^+$ [57]	79.21	81.03	85.40	80.27
DDDNet$^+$ [58]	83.20	86.66	89.12	83.05
IsoSSL-MoCo$^+$ [20]	86.77	88.35	91.44	86.77
ReNeXt3D-101+LSTM$^+$ [59]	84.32	87.60	89.33	82.22
MIGCN$^+$ [82]	80.37	82.33	88.52	82.34
FacialUnits [60]	81.22	89.27	89.77	85.30
2s-STGCN [55]	80.44	86.71	90.04	84.28
VBFLFFA [18]	86.51	92.33	95.21	87.02
ResNet3D$^+$ [83]	**89.37***	**93.78***	94.79	**91.30***
ViViT$^+$ [84]	82.90	90.70	83.44	88.62
VDMoE	87.70	92.91	**95.03***	90.40

Notes: In this table, since there is only the drowsiness estimation task, the VDMoE only applied the $\mathcal{L}_{drow}$ loss.

F. Performance on Other Datasets

To verify the effectiveness of VDMoE in wider scenarios, we assessed it on two other public datasets (i.e., FatigueView and V4V). The results are presented in Table VIII, IX. For the multi-task remote physiological measurement, the VDMoE still significantly outperformed all methods. For the single-task drowsiness estimation in the FatigueView dataset, despite the VDMoE still outperforming most of the baselines, the ResNet3D stands out with the highest accuracy (89.37%), F1-score (93.78%), and specificity (91.30%). It suggests that the ResNet3D is particularly effective at identifying true negatives, i.e., correctly identifying awake states. However, considering the significantly larger computational cost of ResNet3D, we believe that the VDMoE shows competitive single-task performance.

VI. DISCUSSION

Firstly, the VDMoE model's outstanding performance, particularly in comparison with single-task models and multi-task baselines, underscores the effectiveness of its architecture. Notably, the model's superior accuracy in drowsiness estimation, even when compared to specialized single-task models, suggests that a well-designed multi-task model does not necessarily trade off performance in individual tasks for versatility. Besides, the better performance of VDMoE compared to multi-task baselines without task-specific feature subspace construction (e.g., ResNet3D and ViViT) illustrates that shared low-level information extraction and separate task-related feature space can enhance performance across related tasks by leveraging common features. Moreover, the comparative underperformance of models relying solely on facial video inputs versus those utilizing preprocessed facial features highlights the potential limitations of raw video data in capturing subtle physiological or behavioral cues. This finding suggests that preprocessing steps or feature engineering might still play a crucial role in optimizing model performance,

even in an era where end-to-end deep learning models are prevalent. Particularly, considering both facial features at the action level (e.g. blinks and expressions) and changes in color (STMap) during pre-processing may help extract changes in physiological indicators associated with the driver states.

In addition, the inclusion of a prior-driven loss, denoted as $\mathcal{L}_{align}$ plays a pivotal role in aligning the model's outputs with prior knowledge. This loss function helps guide the learning process, which ensures that the model not only optimizes estimation accuracy but also considers the underlying structure or relationships within the data that are known a priori. By incorporating $\mathcal{L}_{align}$, the model is encouraged to learn representations that are consistent with the findings in previous research (i.e., the correlation between drowsiness and cognitive load), potentially leading to more robust and generalizable performance. In the future, exploring more sophisticated forms of prior-driven losses that can be explicitly guided by subtle relationships among tasks could further improve the model's robustness and generalizability.

Lastly, the choice to structure the VDMoE as a multi-task MoE model using MLPs, avoiding more complex networks, is to ensure both the model's performance and its computational efficiency. While 3DCNNs and Transformers have demonstrated remarkable capabilities in capturing spatial and temporal dependencies in data, they come with significantly higher computational costs and parameter counts. For real-time applications, such as driver state monitoring systems where low latency is crucial, the efficiency of the model is as important as its accuracy. MLPs, by contrast, offer a simpler and more computationally efficient alternative. When structured as part of an MoE framework, MLPs can be highly effective in capturing complex relationships within the data. Each "expert" in the MoE model can specialize in different aspects or features of the data, and the "gating" mechanism can learn to dynamically allocate computational resources by activating relevant experts for a given task or input. This allows the VDMoE to maintain a balance between model complexity and computational efficiency, making it suitable for deployment in real-world settings where resources may be constrained. Future studies should investigate the extension of the MoE framework to incorporate other types of network architectures, such as lightweight CNNs or compact Transformer variants, which could offer a pathway to further decrease computational demands in the deployment.

VII. Limitations

While the VDMoE model demonstrates impressive performance in multi-task driver state and physiological measurement estimation in conditional autonomous driving, some limitations still exist in the current study. Firstly, although the VDMoE achieved a balance between performance and computational efficiency through its architecture, the challenges for model interpretability are still unsolved. As a neural network with complex high-dimensional mapping and nonlinear transform, understanding why the model makes specific decisions or how it differentiates between tasks at a granular level is important in building trustworthy AI systems [88].

Secondly, while this study proposed new multi-task datasets in the context of driving, the generalizability of the model to real-world scenarios across different individuals, driving conditions, and sensor setups remains a concern. The potential for biases between the training and testing data could impact the model's effectiveness and fairness when deployed in diverse real-world environments. Particularly, real-world driving conditions are dynamic, with rapid changes in external factors (e.g., light, weather, traffic conditions, occlusions), which might be challenging for assessing the driver's physiological indicators or state with the camera. The model's ability to adapt to these changes in real-time, especially without retraining or manual adjustments, is an area that requires further exploration. Besides, in the MCDD dataset proposed in this paper, limited by the sample size, we are unable to consider the above factors. Meanwhile, there is currently a lack of a multi-task dataset that takes into account multiple challenges in a real environment simultaneously. Therefore, we expect the collection of a more complete dataset and the validation of the models on the dataset as a future work

Lastly, in this paper, the way we label the cognitive load level is a discretization method designed by ourselves based on the NASA-TLX questionnaire. Although this approach demonstrates a certain degree of robustness in terms of performance, it still lacks theoretical support. Therefore, we also expect future work exploring more solid discretization criteria for cognitive load categorization based on subjective questionnaires.

VIII. Conclusion

In this study, we introduced the innovative VDMoE model, tailored for comprehensive driver state and physiological measurement estimation, leveraging the rich cues available in facial videos. Our approach uniquely combines the robustness of a MoE framework with the simplicity and efficiency of MLP, and carefully avoids the computational burdens that are often associated with more complex networks like 3DCNNs and Transformers. The VDMoE model is meticulously designed to distill crucial information from facial features. It employs a prior-driven loss function, $\mathcal{L}_{align}$, to align model predictions with known physiological and behavioral patterns, and enhances both accuracy and generalizability. The incorporation of this loss function, alongside the model's architecture, facilitates a nuanced understanding of both temporal and spatial dynamics in the data, enabling superior performance across tasks of drowsiness, cognitive load, HR, and RR estimation. VDMoE is intended as the core component of a real-time DMS in SAE Level-2/3 driving contexts. Given its lightweight architecture and video-based input design, VDMoE can be directly integrated into an in-vehicle system by processing RGB streams from a cabin-facing camera to infer both cognitive load and drowsiness. The model's multi-task outputs can be fused with vehicle control modules or human-machine interfaces (HMI) to provide adaptive alerts, issue take-over requests, or log physiological states for long-term behavioral analysis. We believe this integration pathway highlights the model's deployment potential in future intelligent cockpit systems.

Moreover, we presented the MCDD, a comprehensive, large-scale dataset for the monitoring of driver states, designed to explore and estimate the multiple states and the physiological indicators within conditional autonomous driving scenarios, utilizing RGB video data. This dataset was collected based on a driving simulator platform, with strict experiment methodologies employed to accurately reflect the co-occurrence of multi-dimensional cognitive load and drowsiness in real-world conditional autonomous driving environments. We hope that the release of MCDD will benefit future research in this area, particularly in addressing practical real-world concerns. In future work, we will continue to design and collect new multi-task datasets and validate our proposed VDMoE in a real-world conditioned autonomous driving environment. In future work, we will continue to design and collect new multi-task datasets and validate our proposed VDMoE in a real-world conditioned autonomous driving environment, especially in more extreme cases, such as when drivers are experiencing very high cognitive load. Moreover, advancements in aligning drowsiness with cognitive workload distribution by integrating human factor knowledge into data-driven approaches are anticipated. We also plan to explore the integration of VDMoE into real-time embedded platforms and its interaction with existing DMS software pipelines as future work.

REFERENCES

[1] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3017–3032, 2015.

[2] Z. Peng, S. Gao, Z. Li, B. Xiao, and Y. Qian, "Vehicle safety improvement through deep learning and mobile sensing," *IEEE network*, vol. 32, no. 4, pp. 28–33, 2018.

[3] S. International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," *SAE international*, vol. 4970, no. 724, pp. 1–5, 2018.

[4] G. Bäumler, "On the validity of the yerkes-dodson law," *Studia Psychologica*, vol. 36, no. 3, p. 205, 1994.

[5] J. R. Perello-March, C. G. Burns, R. Woodman, M. T. Elliott, and S. A. Birrell, "Driver state monitoring: Manipulating reliability expectations in simulated automated driving scenarios," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 6, pp. 5187–5197, 2021.

[6] M. A. Recarte and L. M. Nunes, "Effects of verbal and spatial-imagery tasks on eye fixations while driving." *Journal of experimental psychology: Applied*, vol. 6, no. 1, p. 31, 2000.

[7] E. Muhrer and M. Vollrath, "The effect of visual and cognitive distraction on driver's anticipation in a simulated car following scenario," *Transportation research part F: traffic psychology and behaviour*, vol. 14, no. 6, pp. 555–566, 2011.

[8] S. Soares, T. Monteiro, A. Lobo, A. Couto, L. Cunha, and S. Ferreira, "Analyzing driver drowsiness: From causes to effects," *Sustainability*, vol. 12, no. 5, p. 1971, 2020.

[9] D. He, B. Donmez, C. C. Liu, and K. N. Plataniotis, "High cognitive load assessment in drivers through wireless electroencephalography and the validation of a modified n-back task," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 4, pp. 362–371, 2019.

[10] M. Doudou, A. Bouabdallah, and V. Berge-Cherfaoui, "Driver drowsiness measurement technologies: Current research, market solutions, and challenges," *International Journal of Intelligent Transportation Systems Research*, vol. 18, pp. 297–319, 2020.

[11] Q. Meteier, M. Capallera, S. Ruffieux, L. Angelini, O. Abou Khaled, E. Mugellini, M. Widmer, and A. Sonderegger, "Classification of drivers' workload using physiological signals in conditional automation," *Frontiers in psychology*, vol. 12, p. 596038, 2021.

[12] Y. Qu, H. Hu, J. Liu, Z. Zhang, Y. Li, and X. Ge, "Driver state monitoring technology for conditionally automated vehicles: Review and future prospects," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[13] F. Hu, L. Zhang, X. Yang, and W.-A. Zhang, "Eeg-based driver fatigue detection using spatio-temporal fusion network with brain region partitioning strategy," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[14] D. He, Z. Wang, E. B. Khalil, B. Donmez, G. Qiao, and S. Kumar, "Classification of driver cognitive load: Exploring the benefits of fusing eye-tracking and physiological measures," *Transportation research record*, vol. 2676, no. 10, pp. 670–681, 2022.

[15] A. Picot, S. Charbonnier, and A. Caplier, "On-line detection of drowsiness using brain and visual information," *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans*, vol. 42, no. 3, pp. 764–775, 2011.

[16] G. Li, B.-L. Lee, and W.-Y. Chung, "Smartwatch-based wearable eeg system for driver drowsiness detection," *IEEE Sensors Journal*, vol. 15, no. 12, pp. 7169–7180, 2015.

[17] S. Murugan, J. Selvaraj, and A. Sahayadhas, "Detection and analysis: Driver state with electrocardiogram (ecg)," *Physical and engineering sciences in medicine*, vol. 43, no. 2, pp. 525–537, 2020.

[18] L. Yang, H. Yang, H. Wei, Z. Hu, and C. Lv, "Video-based driver drowsiness detection with optimised utilization of key facial features," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[19] G. Sikander and S. Anwar, "A novel machine vision-based 3d facial action unit identification for fatigue detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 2730–2740, 2020.

[20] L. Mou, C. Zhou, P. Xie, P. Zhao, R. Jain, W. Gao, and B. Yin, "Isotropic self-supervised learning for driver drowsiness detection with attention-based multimodal fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 529–542, 2021.

[21] Y. Peng, H. Deng, G. Xiang, X. Wu, X. Yu, Y. Li, and T. Yu, "A multi-source fusion approach for driver fatigue detection using physiological signals and facial image," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[22] Y. Huang, C. Liu, F. Chang, and Y. Lu, "Self-supervised multi-granularity graph attention network for vision-based driver fatigue detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[23] X.-P. Huynh, S.-M. Park, and Y.-G. Kim, "Detection of driver drowsiness using 3d deep neural network and semi-supervised gradient boosting machine," in *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*. Springer, 2017, pp. 134–145.

[24] G. Du, T. Li, C. Li, P. X. Liu, and D. Li, "Vision-based fatigue driving recognition method integrating heart rate and facial features," *IEEE transactions on intelligent transportation systems*, vol. 22, no. 5, pp. 3089–3100, 2020.

[25] M. Z. Hasan, J. Chen, J. Wang, M. S. Rahman, A. Joshi, S. Velipasalar, C. Hegde, A. Sharma, and S. Sarkar, "Vision-language models can identify distracted driver behavior from naturalistic videos," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[26] S. Ayas, B. Donmez, and X. Tang, "Drowsiness mitigation through driver state monitoring systems: a scoping review," *Human factors*, p. 00187208231208523, 2023.

[27] A. Wang, C. Huang, J. Wang, and D. He, "The association between physiological and eye-tracking metrics and cognitive load in drivers: A meta-analysis," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 104, pp. 474–487, 2024.

[28] J. Wang, H. Lu, A. Wang, X. Yang, Y. Chen, D. He, and K. Wu, "Physmle: Generalizable and priors-inclusive multi-task remote physiological measurement," *arXiv preprint arXiv:2405.06201*, 2024.

[29] Q. Massoz, T. Langohr, C. François, and J. G. Verly, "The ulg multimodality drowsiness database (called drozy) and examples of use," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–7.

[30] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," in *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*. Springer, 2017, pp. 117–133.

[31] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[32] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Springer International Publishing, 2020, pp. 387–405.

[33] C. Yang, Z. Yang, W. Li, and J. See, "Fatigueview: A multi-camera video dataset for vision-based drowsiness detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 233–246, 2022.

[34] W. Jo, R. Wang, G.-E. Cha, S. Sun, R. K. Senthilkumaran, D. Foti, and B.-C. Min, "Mocas: A multimodal dataset for objective cognitive workload assessment on simultaneous tasks," *IEEE Transactions on Affective Computing*, 2024.

[35] H. Yang, J. Wu, Z. Hu, and C. Lv, "Real-time driver cognitive workload recognition: Attention-enabled learning with multimodal information fusion," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 5, pp. 4999–5009, 2023.

[36] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[37] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2016.

[38] D. Hunter, H. Yu, M. S. Pukish III, J. Kolbusz, and B. M. Wilamowski, "Selection of proper neural network sizes and architectures—a comparative study," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 2, pp. 228–240, 2012.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[41] D. J. Saxby, G. Matthews, J. S. Warm, E. M. Hitchcock, and C. Neubauer, "Active and passive fatigue in simulated driving: discriminating styles of workload regulation and their safety impacts." *Journal of experimental psychology: applied*, vol. 19, no. 4, p. 287, 2013.

[42] S. A. Mansi, G. Barone, C. Forzano, I. Pigliautile, M. Ferrara, A. L. Pisello, and M. Arnesano, "Measuring human physiological indices for thermal comfort assessment through wearable devices: A review," *Measurement*, vol. 183, p. 109872, 2021.

[43] M. Deng, A. Gluck, Y. Zhao, D. Li, C. C. Menassa, V. R. Kamat, and J. Brinkley, "An analysis of physiological responses as indicators of driver takeover readiness in conditionally automated driving," *Accident Analysis & Prevention*, vol. 195, p. 107372, 2024.

[44] M. Lohani, B. R. Payne, and D. L. Strayer, "A review of psychophysiological measures to assess cognitive states in real-world driving," *Frontiers in human neuroscience*, vol. 13, p. 57, 2019.

[45] J. Cui, Z. Lan, O. Sourina, and W. Müller-Wittig, "Eeg-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7921–7933, 2022.

[46] A. Alaimo, A. Esposito, C. Orlando, and A. Simoncini, "Aircraft pilots workload analysis: heart rate variability objective measures and nasa-task load index subjective evaluation," *Aerospace*, vol. 7, no. 9, p. 137, 2020.

[47] M. Cella and T. Chalder, "Measuring fatigue in clinical and community settings," *Journal of psychosomatic research*, vol. 69, no. 1, pp. 17–22, 2010.

[48] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[49] W. Wang, A. C. Den Brinker, S. Stuijk, and G. De Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[50] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiological measurement*, vol. 35, no. 5, p. 807, 2014.

[51] H. Lu, H. Han, and S. K. Zhou, "Dual-gan: Joint bvp and noise modeling for remote physiological measurement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 404–12 413.

[52] Z. Yu, Y. Shen, J. Shi, H. Zhao, Y. Cui, J. Zhang, P. Torr, and G. Zhao, "Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1307–1330, 2023.

[53] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 400–19 411, 2020.

[54] G. Narayanswamy, Y. Liu, Y. Yang, C. Ma, X. Liu, D. McDuff, and S. Patel, "Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7914–7924.

[55] J. Bai, W. Yu, Z. Xiao, V. Havyarimana, A. C. Regan, H. Jiang, and L. Jiao, "Two-stream spatial–temporal graph convolutional networks for driver drowsiness detection," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13 821–13 833, 2021.

[56] M. Su-Gang, Z. Chen, S. Han-Lin, and H. Jungang, "Yawning detection algorithm based on convolutional neural network," *Comput. Sci*, 2018.

[57] H. Yang, L. Liu, W. Min, X. Yang, and X. Xiong, "Driver yawning detection based on subtle facial action recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 572–583, 2020.

[58] S. Park, F. Pan, S. Kang, and C. D. Yoo, "Driver drowsiness detection system based on feature representation learning using various deep networks," in *Asian conference on computer vision*. Springer, 2016, pp. 154–164.

[59] L. Zhao, Z. Wang, G. Zhang, and H. Gao, "Driver drowsiness recognition via transferred deep 3d convolutional network and state probability vector," *Multimedia Tools and Applications*, vol. 79, no. 35, pp. 26 683–26 701, 2020.

[60] Q. Cheng, W. Wang, X. Jiang, S. Hou, and Y. Qin, "Assessment of driver mental fatigue using facial landmarks," *IEEE Access*, vol. 7, pp. 150 423–150 434, 2019.

[61] J. Wang, H. Lu, A. Wang, Y. Chen, and D. He, "Hierarchical style-aware domain generalization for remote physiological measurement," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1635–1643, 2024.

[62] J. Wang, X. Wei, H. Lu, Y. Chen, and D. He, "Condiff-rppg: Robust remote physiological measurement to heterogeneous occlusions," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–13, 2024.

[63] A. Das, H. Lu, H. Han, A. Dantcheva, S. Shan, and X. Chen, "Bvpnet: Video-to-bvp signal prediction for remote heart rate estimation," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 01–08.

[64] X. Niu, H. Han, J. Zeng, X. Sun, S. Shan, Y. Huang, S. Yang, and X. Chen, "Automatic engagement prediction with gap feature," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 599–603.

[65] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8583–8595, 2021.

[66] D. Kahneman, *Attention and effort*. Citeseer, 1973, vol. 1063.

[67] G. R. J. Hockey, "Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework," *Biological psychology*, vol. 45, no. 1-3, pp. 73–93, 1997.

[68] G. Merlhiot and M. Bueno, "How drowsiness and distraction can interfere with take-over performance: A systematic and meta-analysis review," *Accident Analysis & Prevention*, vol. 170, p. 106536, 2022.

[69] J. S. Warm, R. Parasuraman, and G. Matthews, "Vigilance requires hard mental work and is stressful," *Human factors*, vol. 50, no. 3, pp. 433–441, 2008.

[70] J. Ma, J. Gu, H. Jia, Z. Yao, and R. Chang, "The relationship between drivers' cognitive fatigue and speed variability during monotonous daytime driving," *Frontiers in psychology*, vol. 9, p. 459, 2018.

[71] D. L. Paulhus, S. Vazire *et al.*, "The self-report method," *Handbook of research methods in personality psychology*, vol. 1, no. 2007, pp. 224–239, 2007.

[72] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[73] Y. Liu and C. D. Wickens, "Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics," *Ergonomics*, vol. 37, no. 11, pp. 1843–1854, 1994.

[74] S. M. Jaeggi, M. Buschkuehl, W. J. Perrig, and B. Meier, "The concurrent validity of the n-back task as a working memory measure," *Memory*, vol. 18, no. 4, pp. 394–412, 2010.

[75] Y. Liang and J. D. Lee, "Combining cognitive and visual distraction: Less than the sum of its parts," *Accident Analysis & Prevention*, vol. 42, no. 3, pp. 881–890, 2010.

[76] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[77] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *International journal of neuroscience*, vol. 52, no. 1-2, pp. 29–37, 1990.
[78] J. I. Campbell and V. A. Thompson, "Morepower 6.0 for anova with relational confidence intervals and bayesian analysis," *Behavior research methods*, vol. 44, pp. 1255–1265, 2012.
[79] C. Ahlström and A. Anund, "Development of sleepiness in professional truck drivers: Real-road testing for driver drowsiness and attention warning (ddaw) system evaluation," *Journal of sleep research*, p. e14259, 2024.
[80] S. Yang, J. Kuo, M. G. Lenné, M. Fitzharris, T. Horberry, K. Blay, D. Wood, C. Mulvihill, and C. Truche, "The impacts of temporal variation and individual differences in driver cognitive workload on ecg-based detection," *Human factors*, vol. 63, no. 5, pp. 772–787, 2021.
[81] A. Revanur, Z. Li, U. A. Ciftci, L. Yin, and L. A. Jeni, "The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2760–2767.
[82] F. Wei, J. Yang, Y. Wang, L. Lin, and H. Zhang, "Prior knowledge-guided multi-information graph convolutional network for driver drowsiness detection," *Expert Systems with Applications*, vol. 275, p. 127028, 2025.
[83] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.
[84] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.
[85] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2019.
[86] J. Wang, A. Wang, H. Hu, K. Wu, and D. He, "Multi-source domain generalization for ecg-based cognitive load estimation: Adversarial invariant and plausible uncertainty learning," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1631–1635.
[87] J. Wang, H. Lu, H. Han, Y. Chen, D. He, and k. Wu, "Generalizable remote physiological measurement via semantic-sheltered alignment and plausible style randomization," *IEEE Transactions on Instrumentation and Measurement*, 2024.
[88] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information fusion*, vol. 99, p. 101805, 2023.

Jiyao Wang received the B.Eng. degree in Software Engineering from Sichuan University, Chengdu, China in 2021, M.Sc. degree in Big Data Technology from the Hong Kong University of Science and Technology (HKUST), Hong Kong S.A.R., China, in 2022, and a Ph.D. degree at HKUST, Guangzhou campus. His research interests include physiological signal measurement, intelligent transport systems, and human factors.

Xiao Yang obtained a bachelor's degree in Computing Science at Sichuan Agricultural University and is currently pursuing master degree at HKUST, Guangzhou campus. His research interests include physiological signal measurement, human factors, and remote sensing.

Zhenyu Wang received the B.S. degree in civil engineering from the University of Jinan and the M.S. degree in energy and power engineering from the Sun Yat-sen University. He is currently pursuing the Ph.D. degree with HKUST(GZ). His current research interests include human factors, thermal comfort, and driving safety.

Ximeng Wei received his bachelor's degree in Refrigeration and Cryogenic engineering from Central South University, China in 2022. For now, he is pursuing the master's degree at the University of Hong Kong, and is the founder of the Hunch Innovation Innovation Technology Co., Ltd, Shenzhen, China.

Ange Wang received his bachelor's degree in traffic engineering from East China Jiaotong University in 2019 and M.Sc. degree in communication and transportation engineering from the Beijing University of Technology in 2022. He is now a Ph.D. student in the Intelligent Transportation Thrust at the Hong Kong University of Science and Technology (Guangzhou).

Dengbo He received his bachelor's degree from Hunan University in 2012, M.S. degree from the Shanghai Jiao Tong University in 2016 and Ph.D. degree from the University of Toronto in 2020. He is currently an assistant professor from the Intelligent Transpiration Trust and Robotics and Autonomous Systems Thrust, the HKUST(Guangzhou). He is also affiliated with the Department of Civil and Environmental Engineering, HKUST, Hong Kong SAR. From 2020 to 2021, he was a post-doctoral fellow at the University of Toronto.

Kaishun Wu (Fellow, IEEE) received the Ph.D. degree in computer science and engineering from HKUST, Hong Kong, in 2011. He was a Distinguished Professor and the Director of Guangdong Provincial Wireless Big Data and Future Network Engineering Center with Shenzhen University, Shenzhen, China. In 2022, he joined HKUST (GZ) as a Full Professor with DSA Thrust and IoT Thrust. He is an Active Researcher with more than 200 papers published on major international academic journals and conferences, as well as more than 100 invention patents, including 12 from the USA. He is an IET, AAIA, and IEEE Fellow.