

# Driving Cognitive Alertness Detecting Using Evoked Multimodal Physiological Signals Based on Uncertain Self-Supervised Learning

Pengbo Zhao, Cheng Lian<sup>id</sup>, *Member, IEEE*, Bingrong Xu<sup>id</sup>, *Member, IEEE*, Yixin Su<sup>id</sup>, and Zhigang Zeng<sup>id</sup>, *Fellow, IEEE*

**Abstract**—Multimodal physiological signals play a pivotal role in drivers' perception of work stress. However, the scarcity of labels and the multitude of modalities render the utilization of physiological signals for driving cognitive alertness detection challenging. We thus propose a multimodal physiological signal detection model based on self-supervised learning. First, in order to mine the intrinsic information of data and enable data to highlight effective information, we introduce a multiscale entropy (MSE) evoked attention mechanism. Secondly, the multimodal patches undergo processing through a novel cascaded attention mechanism. This attention mechanism is rooted in patch-level interactions within each modality, progressively integrating and interacting with other modalities in a cascading manner, thereby mitigating computational complexity. Moreover, a multimodal uncertainty-aware module is devised to effectively cope with intricate variations in the data. This module enhances its generalization ability through the incorporation of uncertain resampling. Experiments were conducted on the DriveDB dataset and the CogPilot dataset with both the linear probing and the fine-tuning evaluation protocols. Experimental results in subject-dependent setting show that our model significantly outperforms previous competitive baselines. In the linear probing evaluation, our model achieves on average 6.26%, 6.64%, and 7.75% improvements in Accuracy (Acc), Recall (Rec), and F1 Score. It also outperforms other models by 7.96% in Acc, 9.13% in Rec, and 9.2% in F1 using the fine-tuning evaluation. Furthermore, our model also demonstrates robust performance in subject-independent setting.

**Index Terms**—Multimodal physiological signals, self-supervised learning, multiscale entropy, multimodal uncertainty-aware, multimodal cascaded attention.

Manuscript received 1 January 2024; revised 24 March 2024 and 23 May 2024; accepted 3 June 2024. Date of publication 7 June 2024; date of current version 13 June 2024. This work was supported by the National Natural Science Foundation of China under Grant 62176193 and Grant 62206204. (*Corresponding author: Cheng Lian.*)

Pengbo Zhao, Cheng Lian, Bingrong Xu, and Yixin Su are with the School of Automation, Wuhan University of Technology, Wuhan 430074, China (e-mail: pengbozhao@whut.edu.cn; chenglian@whut.edu.cn; bingrongxu@whut.edu.cn; suyixin@whut.edu.cn).

Zhigang Zeng is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: zgeng@hust.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2024.3410990

## I. INTRODUCTION

AS DRIVING automation advances and intelligent transportation develops rapidly, vehicle operators, including drivers, pilots and crew members are increasingly required to perform multiple tasks while driving. These demands have progressively outpaced the cognitive capabilities of the vehicle operators [1]. Due to the constraints of human perception and cognition, engaging in high cognitive workloads prompts vehicle operators to concentrate on one stimulus, potentially overlooking other critical tasks or information. This can result in human errors and potentially fatal accidents [2]. Hence, there is a need for comprehensive monitoring of the vehicle operators' cognitive state, timely detection and alleviation of mental fatigue and negative stress emotions, or alerting to potential driving risks. Recent studies demonstrate that vehicle operators' cognitive load can be tracked through various measures, including physical signals, operational signals and behavioral signals [3], [4]. Behavioral signals and operational signals are highly susceptible to external variables. With the advancement of wearable physiological measurement technology and artificial intelligence, physiological signals have emerged as one of the most promising methods for assessing cognitive load. Physiological signal data are the most essential mapping of neural and psychological stress and can provide the most objective measurement of the vehicle operators' cognitive state [5]. Several physiological signals, such as electrocardiography (ECG), electromyography (EMG), electrodermal activity (EDA) and respiration (RESP), contain information that can offer insights into the driver's neurocognitive status. This information can be utilized to provide feedback to driver assistance systems, enabling the delivery of safe driving recommendations [6], [7]. Specifically, we aim to construct reliable and robust algorithms to classify and estimate driver cognitive states through multimodal physiological signals.

Confronted with the dynamically changing scenarios of driving, multimodal physiological signals initiate a cascade of physiological responses, each physiological signal bearing its distinct strengths and limitations. We visualize the multimodal physiological signals of the complete driving task for two

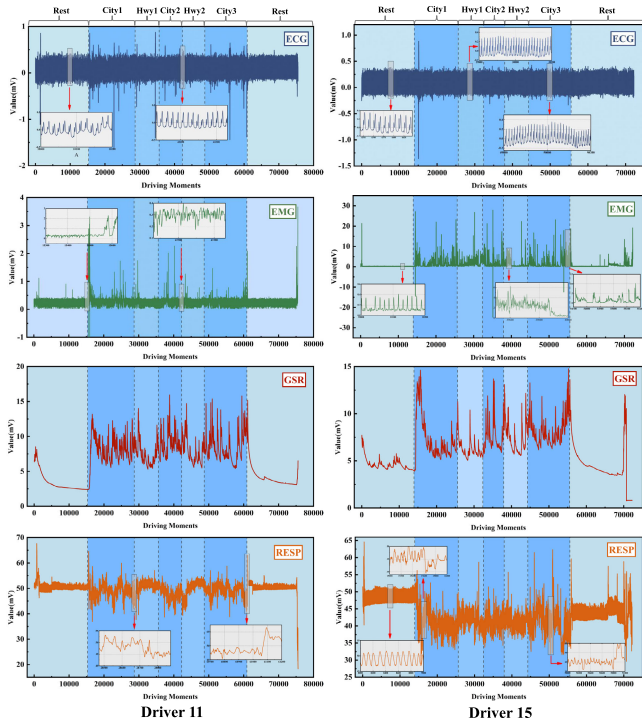


Fig. 1. Comparison of multimodal physiological signals between two subjects on the same complete driving task, with a magnified view applied to certain areas. The entire driving task comprises two Rest sections, three City sections, and two Highway sections.

subjects in the DriverDB dataset [8], as shown in Fig. 1 with a magnified view: 1) the same physiological signals of the two subjects have similar trends overall, but significant differences emerge during the same driving stage; 2) each modality demonstrates varying sensitivity to risk perception throughout the entire driving task; and 3) each modality exhibits irregular fluctuations across different driving stages, particularly evident when transitioning between stages, where the signals display noticeable jumps. The complex changing characteristics of these differences are currently unclear, posing challenges in the utilization of multimodal physiological data. Regardless of how the data mapping changes, the most fundamental changes in the data can be extracted from the perspective of information theory. Through the application of entropy parameters, we can aptly characterize the ordered changes within complex systems, assess trend states and unveil hidden details within the signals. Therefore, we employ MSE to quantify the changes in the complex characteristics of each modality across various time scales. As illustrated in Fig. 2, the entropy values at various scales have obvious peaks in some driving stages, and this phenomenon also appears in various modalities at the same time. This can be interpreted as a collaborative transformation of their global dynamics, elucidating the specific characteristics of the data in different driving states [10], [11].

In this paper, recognizing the challenges posed by the high cost and subjective nature of obtaining physiological data labels, we propose a multimodal self-supervised approach for detecting cognitive load in drivers and pilots. Specifically,

inspired by the steady-state evoked potential (SSVEP) in brain-computer interfaces [12], [13], we devise a method named MSE evoked attention. In the training phase, we adhere to the multimodal self-supervised paradigm [14] with a random masking strategy. To address the challenges posed by multimodal physiological signals, we devise the multimodal uncertainty-aware module. To enhance the integration of multimodal features, we introduce the multimodal cascade attention module. The performance of the model is fully validated by subject-dependent and subject-independent experimental settings. Subject-dependent experiments facilitate a faster and clearer evaluation of models, circumventing the significant complexity and external interference introduced by cross-subjects. Although the subject-independent experimental setting lead to higher experimental costs, it is essential for evaluating the robustness and generalization ability of the model, thus providing a more accurate reflection of its performance in practical applications. In summary, we make the following contributions:

- 1) We leverage MSE to elicit the essence of the data and construct an evoked attention mechanism. This attention highlights the specific values of the samples, offering clear data guidance for the model.
- 2) We propose a multimodal cascade attention module that is alternately connected with the original attention mechanism. This design aims to enhance the fusion of multimodal physiological signals.
- 3) We introduce a multimodal uncertainty-aware representation to mitigate differences between physiological samples and increase the robustness of the model.
- 4) Extensive experiments are conducted using two publicly available datasets. The results of the experiments demonstrate that our model outperforms other models in both driver and pilot categories.

The remainder of this article is organized as follows. Section II presents related work about supervised learning and self-supervised learning of multimodal physiological signals. Section III describes the proposed model in detail. Section IV discusses the experimental design. The performance of the proposed method is evaluated in Section V. In Section VI, we conclude this paper and provide directions for future work.

## II. RELATED WORK

### A. Full-Supervised Learning for Physiological Signals

In recent years, there has been a notable shift in focus towards physiological signal time series analysis, giving rise to a series of robust baseline classification models that leverage convolutional neural network layers [15], [16]. SCINet [17] effectively models complex dynamic time series by employing multiple convolutions to learn effective representations of recursively downsampled subsequences. Xiao et al. [18] proposed a sparsely connected dynamic sparse network, which leverages various receptive fields to train each sparse layer, exploring under-constrained areas and achieving better performance. Vit\_Arjun [19] inherited the standard ViT architecture, the raw 1D EEG signal is directly sliced into different patches along the time dimension and then fed

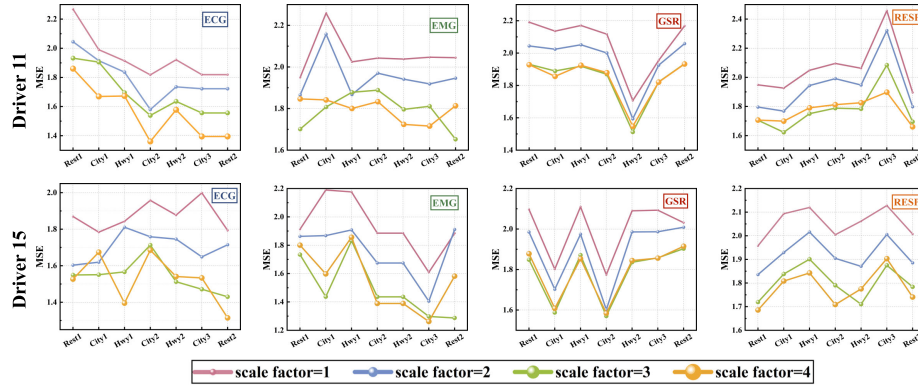


Fig. 2. Comparison results of MSE for multimodal physiological signals from two subjects. The scale factor in MSE is set from 1-4.

into the encoder. Crossformer [20] used dimension-segment-wise to embed the input data into a 2D vector array and then captured cross-time and cross-dimensional dependencies through a two-stage attention layer. In order to capture a wider range of temporal dependencies, Pyraformer [21] proposed a pyramidal attention module, which includes an inter-scale tree structure and an intra-scale neighboring connections model. The aforementioned studies have all focused on the analysis of a single modality, and researchers have gradually expanded to encompass multimodal. Esener [7] proposed a subspace-based feature extraction scheme, using five physiological signals for driver distress recognition. Alonso et al. [22] extracted a set of feature vectors from six biometric signals and applied a combination of principal component analysis and support vector machines for biometric identification, achieving a considerable correct recognition rate. However, most methods heavily depend on labeled physiological data, which can be challenging to obtain in real-world scenarios. Therefore, traditional fully supervised methods may not yield satisfactory results.

### B. Self-Supervised Learning for Physiological Signals

Gradually, people are migrating the self-supervised learning paradigm to physiological signal time series, and the current main work still follows the masked modeling and contrastive learning paradigms. For the first time, TsT [23] used unlabeled multivariate time series data to train a Transformer encoder-based architecture, and proved on multiple data sets that applying the masking modeling paradigm to time series also has huge advantages. PatchTST [24] divided the time series into sub-sequence-level patches, which were then sent to the channel-independent Transformer, which captured local information while benefiting from a longer historical window. Cheng et al. [25] proposed a new time series self-supervised paradigm, TimeMAE, which implemented mask recovery through two pretext tasks: masked codeword classification task and masked representations regression task. Ti-MAE [26] utilized mask modeling as an auxiliary task to train autoencoders with strong representation capabilities at the point level. This resulted in improved performance in prediction and classification tasks. SimMTM [27] also followed the mask modeling self-supervised pre-training

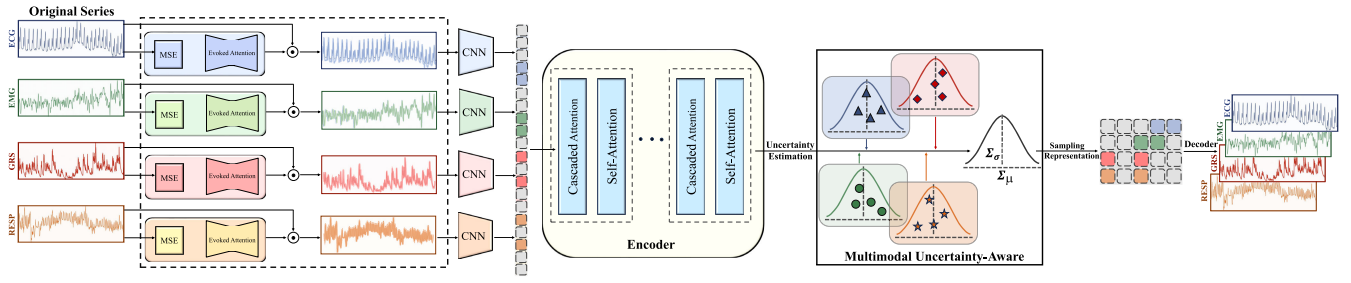
framework, and jointly restored mask time points through multiple other mask sequences to simplify the reconstruction task and reveal the local structure of the manifold. Another self-supervised paradigm is contrastive learning. TS-TCC [28] performed weak enhancement and strong enhancement on the original data, and sent them to the time contrast module and context comparison module respectively to learn robust representation from the time series. These methods are primarily focused on single physiological signals, with limited exploration of self-supervision methods for multimodal physiological signals. Our work aims to provide strong performance by leveraging multimodal physiological signals for pre-training.

## III. METHOD

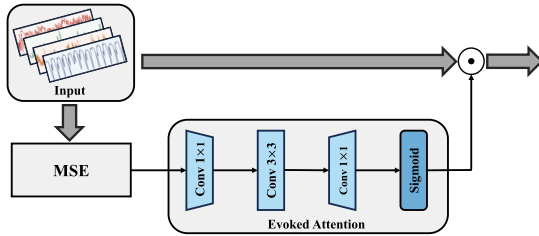
Our proposed framework constitutes a comprehensive multimodal mask self-supervised framework. The decoder is responsible for executing the reconstruction task, thereby completing the learning process of the encoder. Among these components, the MSE evoked attention activates and transforms the original data, the cascaded attention integrates multimodal patches in a cascading manner with low computational complexity, and multimodal uncertainty-aware module enhances the generalization of the model. Fig. 3 shows our proposed method in detail. In this section, we elaborate on the specific structure of each module.

### A. MSE Evoked Attention

Brain-computer interface establishes a direct interactive channel between patients and computers. Recently, there has been an increased focus on the SSVEP signal generated by visual stimulation with a specific flashing frequency. By acquiring and analyzing SSVEP, brain-computer interfaces can convert the patient's intentions into commands for controlling external devices without the need for muscle activity [29]. Inspired by this, we sought to explore potential stimulation methods, considering the complexity of biological systems characterized by variability across various time scales. We adopt MSE as a mediator for the natural activation of physiological signals. MSE characterizes the development and changes of complex systems by measuring entropy across multiple time scales. It quantifies the complexity of



**Fig. 3.** Overview of the proposed self-supervised framework. The input to our model is multimodal physiological signals, including EEG, EMG, EDA and RESP. Each of those inputs is activated using a MSE evoked attention to obtain attention scores. The original input is then pointwise multiplied with the attention score to obtain the transformed data. Next, the transformed data is patched by CNN which backbone is InceptionTime. All patches undergo the masking strategy and are subsequently forwarded to the encoder, which comprises a multi-layer transformer structure. The initial attention mechanism is substituted with Cascaded-Self-Attention, and the Layer Normalization and Feed-Forward layers are omitted for simplicity. The Multimodal Uncertainty-Aware module subsequently conducts uncertainty generalization on the output of the encoder to derive the final representation. Ultimately, the masked patches are reconstructed utilizing the decoder with a single cross-attention layer.



**Fig. 4.** The detailed structure of the MSE evoked attention module.

physiological signals and reveals essential states hidden within the signal. Formally, assume a physiological signal sample,  $\mathbf{U}^c = \{u_1^c, u_2^c, \dots, u_N^c\}$ . First, construct an  $m$  dimensional sequence vector along the time axis as follows:

$$\mathbf{U}_i^c = \{u_i^c, u_{i+1}^c, \dots, u_{i+m-1}^c\}, i = 1, 2, \dots, N-m+1 \quad (1)$$

where  $N$  is the length of the time series, and  $c$  is the number of channels. The distance between two vectors is defined as follows:

$$d[\mathbf{U}_i^c, \mathbf{U}_j^c] = \max \{|u_{i+k}^c - u_{j+k}^c|, k = 0, 1 \dots m-1\} \quad (2)$$

where for each  $1 \leq i, j \leq N-m+1, i \neq j$ , and then count the number  $n$  of  $d$  less than the threshold  $r$ , the probability of lying within the pre-defined threshold as follows:

$$B_i^m(r) = n/(N-m+1) \quad (3)$$

Furthermore, the average of  $B_i^m(r)$  can be obtained by:

$$B_i^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} B_r^m(r) \quad (4)$$

After the dimension of  $m$  is increased by 1, repeat Eqs. (1)-(4) to obtain  $B_i^{m+1}(r)$ . When  $N$  is a finite value, the sample entropy can be expressed as:

$$\text{SampEn}(m, r, N) = -\ln[B^{m+1}(r)/B^m(r)] \quad (5)$$

MSE involves calculating the sample entropy of time series samples at each scale. A coarse-grained new time series as follows:

$$y_j^{(s)} = \frac{1}{s} \sum_{i=(j-1)s+1}^{js} u_i^c \quad (6)$$

where  $s$  is the time scale. The length of the time series after coarse-graining is  $N/s$ . After the coarse-graining process, calculate the sample entropy for each  $y^{(s)}$  and MSE result is obtained.

Unlike the SSVEP signal, which is directly generated by brain electrical activity, MSE, representing an essential change in data, lacks a direct connection with the original physiological signal. Therefore, in order to guide the data to highlight more reliable information, we design an evoked attention module using a simplified inverted bottleneck block, as shown in Fig. 4. Due to the small number of channels in physiological signal data, the overall structure adopts an Expansion-Projection design. Specifically, we calculate the multi-scale entropy of the original data and perform an entropy transformation by multiplying the obtained entropy values with the original data. Then, we apply a convolution operation with a kernel size of 1 to transform the low-dimensional space into a high-dimensional space, setting the expansion coefficient of the channels to 4. Subsequently, a convolution with a kernel size of 3 is applied while keeping the number of channels constant, followed by another convolution with a kernel size of 1 to map back to the low-dimensional space, resulting in an output consistent with the input size. Finally, this output is connected to the original input through a dot product operation. In this way, a new perspective is given to the global awareness of the data to complete operations similar to the brain-computer interface.

## B. Multimodal Uncertainty-Aware Module

Physiological signals provide real-time and sensitive insights into neurological changes induced by the cognitive workload in diverse driving environments. Influenced by an array of subjective factors, physiological signals exhibit considerable variations, particularly in the context of individual distinctions among diverse samples and subjects engaged in the same driving task. These differences cannot be dismissed or eliminated, thus we introduce uncertainty learning into multimodal physiological signals to increase the diversity of samples. During the training process, a batch of data is defined as  $\mathbf{B} = \{(x_i, y_i)\}_{i=1}^B$ , where  $y_i$  indicates the class label of  $x_i$ . When  $\mathbf{B}$  is fed into a module  $\mathcal{M}$ , the output features can

be represented by  $\mathcal{F} = \mathcal{M}(x_i)$ ,  $\mathcal{F} \in \mathbb{R}^{B \times C \times L}$ , the features extracted by the module consist of  $C$  channels, and the length of each channel is  $L$ . The channel-wise feature mean and standard deviation of in  $b$ -th sample can be calculated as:

$$\mu_b = \frac{1}{L} \sum_{l=1}^L \mathcal{F}_l \quad (7)$$

$$\sigma_b^2 = \frac{1}{L} \sum_{l=1}^L (\mathcal{F}_l - \mu_b)^2 \quad (8)$$

assuming that the distribution of each feature statistic follows a multivariate gaussian distribution. We further estimate uncertainty for  $\mu_b$  and  $\sigma_b^2$  over all batch samples:

$$\sum_{\mu}^2 = \frac{1}{B} \sum_{b=1}^B (\mu_b - \mathbb{E}[\mu_b])^2 \quad (9)$$

$$\sum_{\sigma}^2 = \frac{1}{B} \sum_{b=1}^B (\sigma_b - \mathbb{E}[\sigma_b])^2 \quad (10)$$

where  $\mathbb{E}[\cdot]$  is the average of all  $\mu_b$  in a batch.  $\sum_{\mu}$  and  $\sum_{\sigma}^2$  represent the uncertainty estimation of  $\mu_f$  and  $\sigma_f$  in each feature channel [30], [31]. Finally, the reparameterization trick is employed to facilitate the data transformation:

$$\begin{aligned} MUA(f) = & \left( \mathbb{E}[\sigma_b] + \omega_{\sigma} \sum_{\sigma} \right) \left( \frac{\mathcal{F} - \mathbb{E}[\mu_b]}{\mathbb{E}[\sigma_b]} \right) \\ & + \left( \mathbb{E}[\mu_b] + \omega_{\mu} \sum_{\mu} \right) \end{aligned} \quad (11)$$

where  $\omega_{\sigma}$  and  $\omega_{\mu}$  follow the standard gaussian distribution, further modeling uncertainty through random sampling. The resampled feature will be further forwarded to the next module. It is worth noting that this module can only be flexibly applied to various positions during the training phase.

Pre-training on physiological signal data presents unique challenges due to the potential mismatch between pre-training and target, compounded by the unique differences inherent in physiological signals. The intricate nature of physiological signals frequently constrains the effectiveness of pre-training transfer to downstream tasks. As there are no rules for handling complex differences, we start from the data itself to mine the distribution of samples and calculate the uncertainty distribution from it. This approach significantly enhances the model's generalization ability to novel samples within the target domain. The newly sampled instances are derived from a batch size distribution, and the introduction of uncertainty does not yield additional distinctions.

### C. Multimodal Cascaded Attention

The self-attention mechanism facilitates interactions at the patch level, thereby enabling the model to achieve notable performance improvements. However, as the number of patches increases, high computational costs are incurred. In particular, multimodal data are severely limited. With the increase in the number of modalities, employing the pairwise attention interactive fusion strategy will lead to more combinations of data. Consequently, when multiple modal

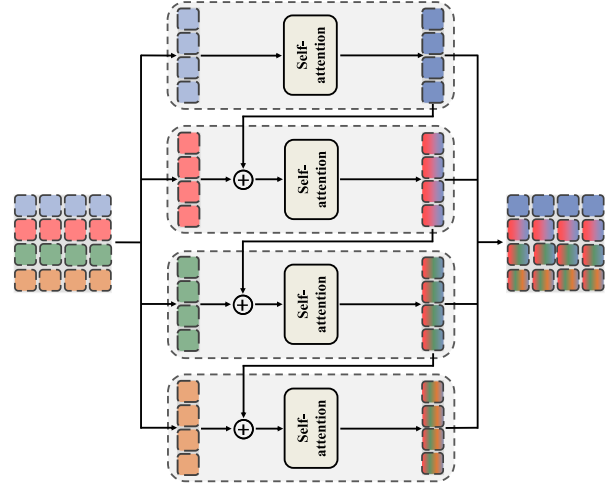


Fig. 5. The detailed structure of cascaded attention module. Four different color patches represent four modalities.

fusions are executed, a substantial amount of redundancy is generated. In order to enhance the efficiency of fusion between multimodal, we introduce the multimodal cascaded attention module, a module that utilizes a cascaded structure to sequentially fuse information from different modalities, as illustrated in Fig. 5. Subsequently, the original attention mechanism is sequentially employed, with the primary objective of augmenting comprehensive interactions among multimodal fusion features.

Specifically, the entirety of multimodal features is categorized into distinct groups based on modality types  $\mathbf{X}_{multi} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$ , where  $N$  represents the number of modalities. At the initial stage, the patches from the first modality are sent to a self-attention mechanism to complete the patch-level interaction within the modality. Formally, the self-attention can be formulated as:

$$\Phi_1^{SA}(\mathbf{X}_1) = \text{Self Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \quad (12)$$

where  $\mathbf{Q}, \mathbf{K}, \mathbf{V}$  are linear projections by multiplying the input patches with the transformation matrix. The sizes of the obtained  $\Phi_1^{SA}(\mathbf{X}_1)$  and  $\mathbf{X}_1$  are consistent. And the length of each modality in the sample is also the same.  $\Phi_1^{SA}(\mathbf{X}_1)$  and the patch of the next modality are directly combined by addition. The formula for calculating fusion is as follows:

$$\mathbf{X}_{12} = \Phi_1^{SA}(\mathbf{X}_1) + \mathbf{X}_2 \quad (13)$$

similarly,  $\mathbf{X}_{12}$  undergoes a self-attention mechanism to interact with dual-modal patches, resulting in the generation of  $\Phi_{12}^{SA}(\mathbf{X}_{12})$ . Following this, the third modality is also integrated with the dual-modal fusion patch  $\Phi_{12}^{SA}(\mathbf{X}_{12})$  through addition to obtain a three-modal fusion  $\mathbf{X}_{123}$ . By analogy,  $\mathbf{X}_{123}$  is sent to self-attention and then merged with the fourth modality. In conclusion, the output from each level constitutes the final output, encompassing single-modal, dual-modal fusion, three-modal fusion and four-modal fusion  $\Phi_i^{CA}(\mathbf{X}_{multi}^D) = \{\Phi_1^{SA}(\mathbf{X}_1), \Phi_2^{SA}(\mathbf{X}_{12}), \Phi_3^{SA}(\mathbf{X}_{123}), \Phi_4^{SA}(\mathbf{X}_{1234})\}$ . This approach diminishes the number of pairwise fusion combinations, enhances the diversity of multimodal fusion, and streamlines the overall model structure.

The resulting output is further subjected to a subsequent self-attention mechanism to accomplish the entirety of the attention operation. According to the standard transformer framework, each attention is followed by the feed-forward layer. The computation can be formulated as:

$$\mathbf{X}_{D+1}^{multi} = \Phi_D^{F'} \left( \Phi_D^{SA} \left( \Phi_D^F \left( \Phi_D^{CA} \left( \mathbf{X}_D^{multi} \right) \right) \right) \right) \quad (14)$$

where  $\Phi_D^F$  and  $\Phi_D^{F'}$  are feed-forward layers, and  $D$  denotes the quantity of attention layers implemented in the model. It is noteworthy that the cascaded attention is executed at the patch level to guarantee interaction among patches. Moreover, this operation is readily extendable to accommodate a greater number of modalities.

In contrast to the original attention mechanism, the cascade attention also yields significant computational complexity. For a modality  $\mathbf{X} \in \mathbb{R}^{n \times d}$  with  $n$  patches of length  $d$  passing through the self-attention mechanism, the computational complexity is expressed as  $O(n^2d)$ , when four modalities are concatenated together and sent to a self-attention mechanism, the computational complexity is  $O(16n^2d)$ . Nevertheless, the cascade attention reduces the computational complexity to  $O(4n^2d)$ , representing a quarter of the original.

#### IV. EXPERIMENTAL DESIGN

##### A. Datasets

1) *DriveDB*: DriveDB [8] is derived from the stress recognition in automobile drivers database. This dataset comprises comprehensive physiological recordings obtained from 16 drivers during real-world open-road driving scenarios. To ensure uniform and consistent processing of all physiological signals, we performed downsampling, reducing the sampling rate of all signals to a common rate of 15.5Hz. Each driver participated in identical and consecutive driving tasks, encompassing Rest, City, and Highway driving conditions. According to the MIT Media Laboratory rating criteria, category labels were assigned to define the stress levels experienced by car drivers under different traffic conditions. Specifically, the Resting phase (comprising Initial Rest and Final Rest) of the driving segment was categorized as Low-Stress. The Highway phase (Highway 1 and Highway 2) was classified as Neutral-Stress, while driving within city environments was labeled as High-Stress. For our analysis, we focused on the physiological data of 10 drivers from the dataset, as this subset included complete ECG, EMG, GSR, and RESP signals, the number of channels for each modality is 1. To suppress unreasonable samples during pressure accumulation and pressure transitions, we began sampling five minutes after the start of each driving segment. Each sample had a non-overlapping duration of 10 seconds and we assigned corresponding labels to each segment.

2) *CogPilot*: CogPilot [32], a dataset focusing on multi-modal physiological monitoring during virtual reality driving tasks. This dataset was meticulously gathered while participants engaged in a series of virtual flight tasks, each characterized by varying levels of difficulty. The participants encountered four distinct levels of difficulty, with these variations introduced through adjustments in wind speed,

turbulence, and visibility conditions. To maintain consistency, we opted to utilize only ECG, EMG, GSR and RESP for our research with channel numbers of 3, 5, 2, and 1 respectively. Furthermore, we standardized the sampling rate for all modalities to 15.5Hz by downsampling from their original rates. Our analysis included data from a total of 30 subjects in the dataset. Each subject actively participated in the experiment in all scenarios. The United States Air Force and MIT Artificial Intelligence Accelerator divide stress labels into four levels based on the difficulty of scenarios, including three different active tasks and one resting state. We conducted non-overlapping 10 seconds segmentations for each type of scene experiment.

##### B. Implementation Details, Baseline and Evaluation Metrics

In this paper, all the experiments are implemented in PyTorch and conducted on a single NVIDIA 3080Ti 12GB GPU. In the pre-training stage, the batch size of all models is set to 128 and the training epochs are configured to 200. The AdamW optimizer is employed with a base learning rate of 1e-4 and weight decay of 0.05. We warm up training for 40 epochs, starting from learning rate 1e-6, and decay it to 0 throughout training using cosine decay. Different from the training stage, in the fine-tuning stage, the learning rate is increased to 1e-3. Warm up epochs are reduced to 5 and training epochs are set to 100.

To conduct a comprehensive evaluation, we compare our method with 10 baseline methods. In full-supervised learning, InceptionTime [16], SCINet [17] and DSN [18] are models based on CNN, Vit\_Arjun [19], Crossformer [20] and Ours (Random Init) is implemented based on Transformer. Random Init refers to the utilization of randomly initialized weights, initiating the training process from scratch. In self-supervised learning, TimeMAE [25] and MultiMAE [14] employ the mask paradigm, and TS-TCC [28], TF-C [33], TS2Vec [34] and CPC [35] are derived from the contrastive learning paradigm.

In all our experiments, we demonstrate the efficacy of the pre-trained model through two widely adopted evaluation protocols: linear probing and fine-tuning evaluation. The linear probing evaluation freezes all parameters of the entire model and exclusively updates the weights of the final classification layer to adapt to downstream tasks. The fine-tuning evaluation adapts all parameters of the model in accordance with downstream tasks without enforcing freezing operations. We utilize the Accuracy(Acc), Recall(Rec) and F1 score as metrics to assess the model's performance in classification tasks. The best results are highlighted in boldface. Reported experimental results are mean and standard deviation values across five independent trials.

#### V. EXPERIMENTAL RESULTS

##### A. Comparison Among Previous Methods

In this section, the performance of the model is thoroughly evaluated through subject-dependent setting, where both the training and test sets include samples from all subjects.

TABLE I  
COMPARISON AMONG OURS MODEL AND BASELINE METHODS

Training strategy	Methods	DriveDB			CogPilot		
		Acc	Rec	F1	Acc	Rec	F1
Full-supervised	InceptionTime [16]	82.69±0.34	81.33±0.37	81.33±0.42	89.17±0.31	89.46±0.27	89.52±0.28
	ViT_Arjun [19]	77.56±0.67	77.68±0.76	75.87±0.65	78.13±0.65	78.78±0.58	78.29±0.53
	Crossformer [20]	79.06±0.45	78.11±0.41	78.19±0.37	83.92±0.38	84.36±0.44	84.49±0.34
	Scinet [17]	75.64±0.32	74.22±0.29	74.15±0.35	77.42±0.16	78.01±0.26	77.84±0.21
	DSN [18]	81.84±0.27	80.71±0.36	80.98±0.31	86.42±0.44	86.77±0.47	87.01±0.35
	Ours (Random.Init)	83.97±0.39	82.79±0.37	82.93±0.33	84.76±0.43	85.56±0.38	85.66±0.47
Full-supervised: Fine-tuning	MultiMAE [14]	79.98±0.42	80.61±0.39	80.20±0.43	75.35±0.41	76.89±0.44	76.72±0.39
	TF-C [33]	83.33±0.10	82.68±0.14	82.79±0.11	85.42±0.15	85.63±0.15	86.12±0.11
	TimeMAE [25]	75.21±0.58	71.78±0.65	71.95±0.52	83.68±0.58	84.20±0.53	84.35±0.52
	TS-TCC [28]	81.83±0.23	80.53±0.25	80.22±0.28	74.10±0.22	75.06±0.25	75.34±0.26
	TS2Vec [34]	78.63±0.37	76.94±0.33	77.15±0.37	85.17±0.36	85.36±0.35	85.69±0.38
	CPC [35]	79.27±0.46	77.65±0.38	77.31±0.48	91.92±0.48	91.93±0.34	91.82±0.42
	Ours	<b>88.46±0.41</b>	<b>87.27±0.35</b>	<b>87.59±0.28</b>	<b>92.00±0.22</b>	<b>92.44±0.39</b>	<b>92.43±0.36</b>
	Ours	<b>88.46±0.41</b>	<b>87.27±0.35</b>	<b>87.59±0.28</b>	<b>92.00±0.22</b>	<b>92.44±0.39</b>	<b>92.43±0.36</b>
Full-supervised: Linear probing	MultiMAE [14]	60.89±0.69	59.51±0.58	58.19±0.57	62.13±0.67	63.48±0.55	62.06±0.49
	TF-C [33]	63.67±0.32	63.87±0.41	64.15±0.19	74.93±0.35	75.49±0.35	75.19±0.34
	TimeMAE [25]	58.97±0.78	57.68±0.56	50.38±0.61	74.02±0.58	74.70±0.64	74.41±0.55
	TS-TCC [28]	60.68±0.43	61.44±0.28	59.55±0.39	61.62±0.45	61.94±0.48	61.92±0.43
	TS2Vec [34]	60.68±0.31	62.36±0.25	59.03±0.33	71.60±0.32	72.32±0.31	72.32±0.31
	CPC [35]	61.47±0.35	60.89±0.42	60.45±0.34	78.41±0.33	78.93±0.37	78.86±0.35
	Ours	<b>65.59±0.33</b>	<b>65.10±0.47</b>	<b>64.72±0.34</b>	<b>79.18±0.42</b>	<b>80.30±0.51</b>	<b>80.21±0.39</b>
	Ours	<b>65.59±0.33</b>	<b>65.10±0.47</b>	<b>64.72±0.34</b>	<b>79.18±0.42</b>	<b>80.30±0.51</b>	<b>80.21±0.39</b>

Specifically, the entire experiment is divided into two training strategies: fully-supervised and self-supervised. The datasets are divided into  $\mathcal{D}_{train}$ :  $\mathcal{D}_{test} = 8$ : 2. We first conduct self-supervised pre-training on the  $\mathcal{D}_{train}$  without using labels, followed by the fine-tuning phase on the  $\mathcal{D}_{train}$ . Finally, the model's performance is evaluated on the  $\mathcal{D}_{test}$ . To ensure consistent data distribution between the training set and the test set, we use stratified sampling on the divided dataset. The experimental results are presented in Table I. All fully supervised methods are trained from scratch using the  $\mathcal{D}_{train}$  without self-supervised pre-training. In comparison to other fully supervised methods, our fully supervised method has certain competitiveness. However, it is imperative to acknowledge that it has yet to attain the best results on CogPilot. In the fine-tuning evaluation, our model demonstrates substantial performance enhancements compared to various fully supervised methods, especially for Ours(Random Init). Overall, in the fine-tuning evaluation, our model achieves on average 6.26%, 6.64%, and 7.75% improvements in Acc, Rec, and F1. On the DriveDB dataset, two self-supervised comparative learning methods, TF-C and TS-TCC, yielded suboptimal results. Within the CogPilot dataset, CPC achieves higher results, which are less than 1% away from the best result. Additionally, various other methods demonstrate commendable performance in this dataset. The results of the linear probing are understandably inferior to that of the fine-tuning evaluation. This is a rational expectation since the linear probing evaluation only adjusts the classification head. Overall, in the linear probing evaluation, our model achieves on average 7.96%, 9.13%, and 9.2% improvements in Acc, Rec, and F1. On the DriveDB dataset, TF-C produces less than 1% of the best performance. It is noteworthy that, exclusively employing Transformer-based structures, MultiMAE and TimeMAE exhibit less than satisfactory performance.

TABLE II  
COMPARISON FOR DIFFERENT PROPORTIONS OF MASK RATIO IN THE DRIVEDB DATASET

Mask ratio	Linear probing			Fine-tuning		
	Acc	Rec	F1	Acc	Rec	F1
5%	63.52±0.53	63.83±0.62	63.08±0.54	86.53±0.45	85.56±0.62	85.70±0.48
15%	65.59±0.35	63.98±0.48	64.23±0.69	88.03±0.45	86.79±0.42	87.18±0.39
25%	64.53±0.34	62.65±0.47	63.02±0.39	87.39±0.33	86.30±0.46	86.52±0.10
35%	65.38±0.53	63.70±0.55	64.15±0.52	86.53±0.53	84.51±0.52	85.18±0.55
45%	63.03±0.12	62.85±0.24	63.18±0.19	85.68±0.12	85.50±0.28	85.02±0.22
55%	65.45±0.38	64.18±0.48	64.68±0.29	87.60±0.24	86.40±0.26	86.77±0.42
65%	<b>65.59±0.33</b>	<b>65.10±0.47</b>	<b>64.72±0.34</b>	<b>88.46±0.41</b>	<b>87.27±0.35</b>	<b>87.59±0.28</b>
75%	64.74±0.31	63.42±0.51	64.29±0.41	87.60±0.14	86.24±0.43	86.66±0.28
85%	64.95±0.42	63.66±0.54	64.15±0.34	85.68±0.38	84.85±0.37	84.70±0.43

TABLE III  
COMPARISON FOR DIFFERENT PROPORTIONS OF MASK RATIO IN THE COGPLOT DATASET

Mask ratio	Linear probing			Fine-tuning		
	Acc	Rec	F1	Acc	Rec	F1
5%	75.93±0.45	76.97±0.49	77.13±0.62	89.42±0.25	90.39±0.38	90.41±0.31
15%	78.26±0.29	78.96±0.35	79.23±0.18	90.09±0.34	90.64±0.28	90.69±0.31
25%	73.77±0.48	74.79±0.39	74.75±0.53	91.00±0.49	91.48±0.35	91.51±0.33
35%	71.19±0.26	71.82±0.34	72.02±0.48	90.50±0.81	91.03±0.55	91.08±0.41
45%	75.60±0.31	76.00±0.64	76.48±0.67	91.67±0.48	92.17±0.45	92.13±0.66
55%	<b>79.18±0.42</b>	<b>80.30±0.51</b>	<b>80.21±0.39</b>	92.00±0.34	92.44±0.45	92.43±0.38
65%	74.85±0.61	76.17±0.42	75.51±0.48	92.07±0.21	92.45±0.37	92.44±0.22
75%	75.18±0.37	76.03±0.51	76.19±0.27	<b>92.25±0.22</b>	<b>92.64±0.39</b>	<b>92.61±0.36</b>
85%	74.52±0.49	75.29±0.67	75.50±0.35	90.08±0.54	90.47±0.55	90.57±0.46

### B. The Influence of the Masking Ratio

The mask self-supervision paradigm involves selectively masking a specific proportion of the input patch. Determining an appropriate mask ratio during the training phase is crucial for enhancing the capabilities of the encoder. Experiments are conducted by adjusting only the masking rate, which ranges from 10% to 90%, as indicated in Tables II and III. On the DriveDB dataset, the best results for both the linear probing and the fine-tuning evaluation are achieved with a mask rate of 65%. As the mask rate decreases, the model performs worse. However, it attained the second-best result at a 15% mask rate, a small number of masks coupled

TABLE IV  
PERFORMANCE ANALYSIS FOR DIFFERENT PROPORTIONS OF TRAINING SET IN THE DRIVERDB DATASET

Training Ratio	Linear probing			Fine-tuning		
	Acc	Rec	F1	Acc	Rec	F1
	10%	57.26±0.28	52.69±0.42	51.28±0.35	67.73±0.46	65.52±0.36
20%	60.04±0.27	55.29±0.53	53.43±0.51	76.92±0.62	75.33±0.37	75.66±0.43
40%	65.17±0.26	59.02±0.19	54.78±0.26	83.12±0.57	81.66±0.36	82.07±0.32
60%	64.74±0.19	60.01±0.29	58.91±0.37	86.32±0.25	84.74±0.64	85.13±0.39
80%	65.23±0.13	63.29±0.46	62.75±0.43	86.96±0.38	85.92±0.56	86.00±0.33
100%	<b>65.59±0.33</b>	<b>65.10±0.47</b>	<b>64.72±0.34</b>	<b>88.46±0.41</b>	<b>87.27±0.35</b>	<b>87.59±0.28</b>

TABLE V  
PERFORMANCE ANALYSIS FOR DIFFERENT PROPORTIONS OF TRAINING SET IN THE COGPILOT DATASET

Training ratio	Linear probing			Fine-tuning		
	Acc	Rec	F1	Acc	Rec	F1
	10%	57.11±0.37	57.54±0.41	58.25±0.43	60.88±0.67	61.62±0.57
20%	69.52±0.55	70.81±0.37	69.77±0.35	78.78±0.29	79.09±0.38	78.88±0.47
40%	70.85±0.22	71.39±0.39	71.89±0.17	87.26±0.24	87.76±0.57	87.79±0.31
60%	71.77±0.12	72.89±0.17	73.01±0.23	89.59±0.33	90.21±0.28	90.16±0.15
80%	74.35±0.49	75.26±0.46	74.52±0.63	90.84±0.41	91.28±0.43	91.36±0.47
100%	<b>79.18±0.42</b>	<b>80.30±0.51</b>	<b>80.21±0.39</b>	<b>92.00±0.22</b>	<b>92.44±0.39</b>	<b>92.43±0.36</b>

with a specific quantity of patches enabled the model to extract more informative features. At high masking rates, specifically above 65%, the model’s performance progressively diminishes. The CogPilot dataset exhibits distinct optimal masking rates for the two evaluation methods, resulting in the attainment of the best results. In the fine-tuning evaluation, the model’s overall performance exhibits an incremental trend with increasing mask rates, reaching a peak at 75%. In the linear probing evaluation, optimal results are attained at a mask rate of 55%. In general, a larger masking ratio poses a sufficiently challenging task, fostering enhanced representational capabilities in the model during the recovery process, and significant disparities between datasets necessitate models to employ varying masking rates.

### C. Tuning With Different Proportions of Training Data

The self-supervised pre-trained model necessitates heightened generalization capabilities, particularly when confronted with sparsely labeled datasets. Therefore, we undertake an investigation into the model’s efficacy by varying the proportion of  $\mathcal{D}_{train}$  during the fine-tuning phase. In the experimental protocol, the proportion of  $\mathcal{D}_{train}$  is adjusted across intervals of (10%, 20%, 40%, 60%, 80%, 100%), while keeping all other parameters constant. The experimental results are shown in Tables IV and V. Consistently, as the proportion of  $\mathcal{D}_{train}$  increases, the performance of the model continues to improve. When the ratio of  $\mathcal{D}_{train}$  is only 10% and 20%, a notable deviation in the model’s performance deviates significantly from the best outcome. When utilizing 40% of  $\mathcal{D}_{train}$ , the model attains a better performance, with a deviation of only 5% from the best result. In the face of regions characterized by limited sample sizes and sparse labels, the construction of a robust self-supervised pre-training model assumes paramount importance.

### D. Analysis About Model Size

In the realm of self-supervision, the potential of the encoder arises not only from its innovative structure but also from

TABLE VI  
MODEL ARCHITECTURE ABLATIONS. W/O MEANS WITHOUT

Method	DriveDB			CogPilot		
	Acc	Rec	F1	Acc	Rec	F1
	Linear probing					
w/o MSE Evoked Attention	63.01±0.76	63.01±0.65	63.57±0.62	73.10±0.43	74.06±0.61	73.61±0.53
w/o Cascaded Attention	65.38±0.56	63.37±0.42	63.54±0.47	73.52±0.42	73.56±0.41	73.88±0.39
w/o Uncertainty-Aware	61.53±0.31	61.97±0.59	60.00±0.41	71.35±0.24	72.53±0.39	72.93±0.18
Ours	<b>65.59±0.33</b>	<b>65.10±0.47</b>	<b>64.72±0.34</b>	<b>79.18±0.42</b>	<b>80.30±0.51</b>	<b>80.21±0.39</b>
Fine-tuning						
w/o MSE Evoked Attention	86.11±0.72	84.23±0.62	84.90±0.70	90.75±0.55	91.27±0.49	91.28±0.67
w/o Cascaded Attention	85.25±0.53	83.90±0.61	84.11±0.54	91.77±0.46	92.24±0.31	92.21±0.23
w/o Uncertainty-Aware	84.40±0.55	82.60±0.33	82.98±0.58	82.26±0.47	87.05±0.49	87.19±0.35
Ours	<b>88.46±0.41</b>	<b>87.27±0.35</b>	<b>87.59±0.28</b>	<b>92.00±0.22</b>	<b>92.44±0.39</b>	<b>92.43±0.36</b>

its size. We conduct experiments aimed at investigating the influence of encoders of varying sizes on the model. We employ incremental adjustments to frequently used parameters, namely  $ly$  (number of layers),  $eb$  (embedding size), and  $ep$  (number of epochs). The controlled ranges are set to (8, 12), (64, 128, 256) and (200, 300) for  $ly$ ,  $eb$  and  $ep$ . We configure a total of 6 different sets of parameters, and the experimental results are shown in Fig. 6. In the linear probing evaluation, with the expansion of encoder dimensions, there is a discernible enhancement in model performance. Simultaneously, this progression is accompanied by a concomitant necessity for an increased number of training epochs. Intriguingly, even with a reduction in dimensions ( $ly$  and  $eb$  set to 8 and 128, respectively), superior results are attainable, notably evidenced in Fig. 6 (c). In contrast, the larger model demonstrates substantial performance gains when assessed on the DriveDB dataset, a phenomenon that may be attributed to the comparatively diminished sample size within this dataset. Throughout the fine-tuning evaluation, with the exception of (8, 64, 200) and (8, 64, 300), the model’s performance exhibited notable proximity across various other parameter configurations. In light of the aforementioned observations, a judicious balance between model performance and computational efficiency led us to the definitive selection of the parameter configuration (8, 128, 200).

### E. Ablation Studies and Parameter Analysis

1) *Influence of Each Component Module*: We conduct ablation studies to investigate the contribution of each essential component of our model. The results are shown in Table VI. When the MSE evoked attention is ablated, multimodal physiological signals are fed directly into the CNN. For each metric in the fine-tuning evaluation, the minimum reduction in the DriveDB dataset is 2%, and the reduction in the CogPilot dataset is generally less than 2%. Surprisingly, the experimental results drop even more dramatically in the linear probing evaluation. The maximum reduction is only 2% in the DriveDB dataset, and the minimum reduction in the CogPilot dataset is 6%. These results indicate that the activation and transformation data are important for the model. Furthermore, when cascade attention is ablated, the attention is set to the original self-attention. Each metric decreases slightly in the fine-tuning evaluation. However, a notable discrepancy of up to 6% is observed in the linear probing evaluation, specifically for the CogPilot dataset. This shows that the cascaded attention improves the performance of the model. Additionally, due consideration is warranted for the strategy



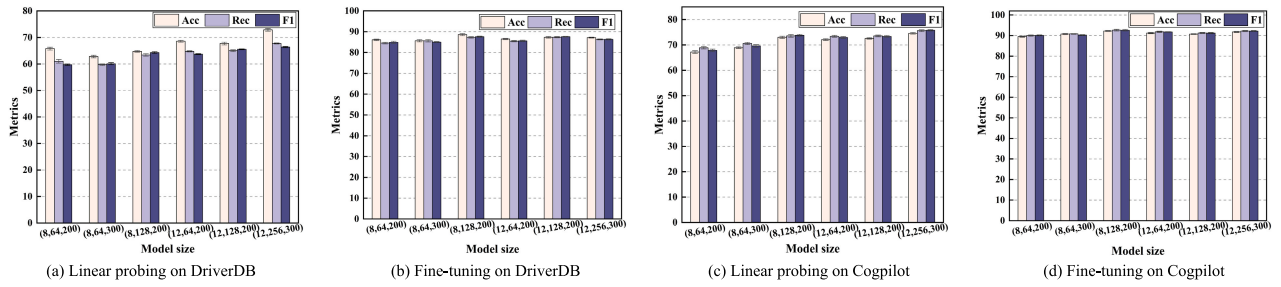


Fig. 6. Performance of encoder with different sizes ( $ly$ ,  $eb$ ,  $ep$ ). Here,  $ly$  denotes the number of layers in the Transformer within the encoder,  $eb$  represents the embedding size, and  $ep$  denotes the number of training epochs.

TABLE VII  
PERFORMANCE OF DIFFERENT INSERTED POSITIONS.  
W/O MEANS WITHOUT

Evaluate Pattern	Position	DriverDB			CogPilot		
		Acc	Rec	F1	Acc	Rec	F1
Linear probing	w/o Uncertainty-Aware	61.53±0.31	61.97±0.59	60.00±0.41	71.35±0.24	72.53±0.39	72.93±0.18
	After CNN	62.82±0.71	62.15±0.65	60.37±0.63	72.35±0.35	73.06±0.42	73.23±0.65
	After Embedding	64.10±0.47	64.44±0.59	63.58±0.87	70.60±0.62	71.62±0.46	71.87±0.46
	After Encoder	62.60±0.77	62.09±0.42	60.78±0.58	71.10±0.31	72.48±0.24	72.15±0.36
	All	<b>65.59±0.33</b>	<b>65.10±0.47</b>	<b>64.72±0.34</b>	<b>79.18±0.42</b>	<b>80.30±0.51</b>	<b>80.21±0.39</b>
Fine-tuning	w/o Uncertainty-Aware	84.40±0.55	82.60±0.33	82.98±0.58	86.26±0.47	87.05±0.49	87.19±0.35
	After CNN	84.61±0.19	83.63±0.27	83.68±0.32	90.34±0.28	90.94±0.29	90.90±0.35
	After Embedding	86.53±0.41	85.04±0.61	85.39±0.68	90.25±0.51	90.72±0.24	90.78±0.38
	After Encoder	83.33±0.46	83.16±0.16	83.22±0.44	86.34±0.23	87.16±0.48	87.15±0.25
	All	<b>88.46±0.41</b>	<b>87.27±0.35</b>	<b>87.59±0.28</b>	<b>92.00±0.22</b>	<b>92.44±0.39</b>	<b>92.43±0.36</b>

of attention stacking. A detailed exploration of this strategy is presented in section V-E. 3). Finally, the removal of the multimodal uncertainty-aware module results in the model achieving its poorest performance, with a reduction of at least 4%. Particularly noteworthy, the difference observed in the linear probing evaluation for the CogPilot dataset is a remarkable 8%. The observed phenomena may be ascribed to the insertion of this module at diverse locations, and a comprehensive exploration of these effects will be discussed in detail in the next section. In summary, each module is essential for the model.

### 2) Influence of The Multimodal Uncertainty-Aware Module:

The Multimodal uncertainty-aware module is deemed a plug-and-play module that can be easily integrated at any position. Table VII presents the experimental results obtained from the two datasets. Here, we integrate this module after each module, namely CNN, Embedding and Encoder. After the insertion of this module at various positions, the majority exhibit improvements when compared to the original module without the uncertainty-aware module. Only a limited number of positions demonstrate performance close to or inferior to that of the original model after insertion. In the DriveDB dataset and the CogPilot dataset, the model demonstrates the most significant improvement after the insertion of uncertainty-aware module into the Embedding and CNN, respectively. Overall, the performance improvement of the model after the uncertainty-aware module is inserted into the Encoder is not that obvious. This phenomenon may be because the uncertainty changes closest to the output layer and the model is no longer training, which exerts a direct impact on the results.

3) Influence of the Multimodal Cascaded Attention: To substantiate the efficacy of the cascade structure, we also investigated various strategies for stacking attention. Specifically, we assume that the encoder consists of  $D$  layers. The

TABLE VIII  
PERFORMANCE OF THE MULTIMODAL CASCADED ATTENTION

Evaluate Pattern	Stacking Strategy	DriverDB			CogPilot		
		Acc	Rec	F1	Acc	Rec	F1
Linear probing	Self-Cascaded	62.82±0.29	61.42±0.43	60.60±0.45	70.02±0.65	70.31±0.39	71.23±0.15
	All-Self	65.38±0.32	63.37±0.48	63.54±0.26	73.52±0.57	73.56±0.52	73.88±0.23
	All-Cascaded	65.45±0.18	62.07±0.23	62.09±0.35	74.52±0.41	75.78±0.55	75.84±0.85
	Alternate	<b>65.59±0.33</b>	<b>65.10±0.47</b>	<b>64.72±0.34</b>	<b>79.18±0.42</b>	<b>80.30±0.51</b>	<b>80.21±0.39</b>
Fine-tuning	Self-Cascaded	85.68±0.14	83.81±0.41	84.27±0.26	91.75±0.41	92.19±0.42	92.17±0.18
	All-Self	85.25±0.44	83.90±0.43	84.11±0.32	91.77±0.46	92.24±0.13	92.21±0.21
	All-Cascaded	86.75±0.75	85.07±0.61	85.63±0.34	90.09±0.39	90.59±0.13	90.56±0.45
	Alternate	<b>88.46±0.41</b>	<b>87.27±0.35</b>	<b>87.59±0.28</b>	<b>92.00±0.22</b>	<b>92.44±0.39</b>	<b>92.43±0.36</b>

Self-Cascaded is defined such that the first  $D/2$  layer employs the cascaded attention, while the last  $D/2$  layer utilizes self-attention. The designations All-Self and All-Cascaded denote configurations where all  $D$  layers are exclusively the self-attention or the cascaded attention, respectively. The Alternate refers to the alternating stacking strategy: (Self + Cascaded)  $\times D/2$ . Table VIII presents the experimental results obtained from the two datasets. Relying on a single attention mechanism fails to yield optimal results. The DriveDB dataset demonstrates a preference for cascaded attention, emphasizing its focus on inter-modality fusion. Conversely, the CogPilot dataset prioritizes intra-modality interactions and prefers using the self-attention. In the linear probing evaluation, Self-Cascaded achieved the worst results on both datasets. In the fine-tuning evaluation, the alternating method achieves the best results. The alternating attention stacking strategy exhibits enhanced adaptability to various datasets. The inclusion of cascaded attention ensures effective fusion among multimodal patches, while the self-attention further enhances the interaction within multimodal fusion patches.

4) Influence of MSE Scales Factor: As the primary tool for stimulating physiological signals, MSE necessitates detailed experimental research on the critical parameter—scale factors. Fig. 7 depicts the results for the two datasets with scale factors in the range of 1 to 10. Collectively, as the scale factor increases, the model's performance exhibits gradual improvement, reaching its best results at a scale factor of 4. As the scale factor continues to increase, the model's performance experiences a gradual decline, eventually stabilizing with minor fluctuations. In the linear probing evaluation, as the scale factor continues to increase, the model's performance undergoes a sharp decline until it stabilizes, suggesting that the evaluation method of frozen parameters is more sensitive to changes in the scale factor. Crucially, we also found that when the scale factor is 8,

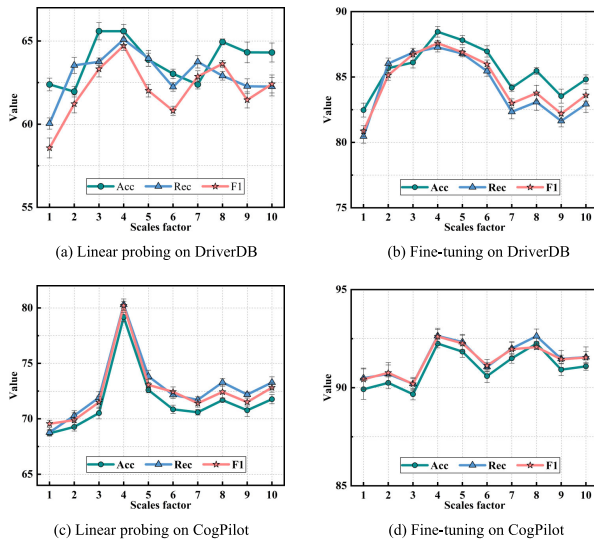


Fig. 7. Effects of the scales factor of MSE on both datasets. Two evaluation protocols are used, with scales factor ranging from 1-10.

the model still has a better performance. As depicted in Figs. 7 (a) and (d), various evaluation metrics approach optimal results. While noticeable changes are not evident in Figs. 7 (b) and (c), a discernible fluctuation in performance improvement is observable. As the scale factor increases, the computation time also exhibits an increment. Ultimately, we opt for a scale factor of 4.

5) *Visualization of Reconstruction Effects*: The results of the reconstruction task are depicted in Fig. 8. We visualized all four physiological signals of a sample. The pre-training task is conducted on the DriveDB dataset, with the mask rate set at 65%. We observe that when confronted with this more challenging reconstruction task, our model excels in reconstructing multimodal physiological signals, thereby contributing to the enhancement of the encoder’s representation learning. Specifically, while accurately reconstructing the overall trends of each modality, significant disparities emerge in certain details, such as fluctuations in amplitude and localized trend offsets.

#### F. Subject-Independent Experiments

To better reflect real-world scenarios, we further explore a more challenging task: subject-independent. In subject-independent experiments, one subject from the dataset is selected as the test subject in turn for testing the model, and the rest subjects are used for training the model. In the pre-training stage, only the rest subjects’ data are used to train the model. The experimental results are presented in Table IX. This table displays the classification accuracy for all 10 subjects (S1-S10) in the DriveDB dataset. Additionally, we calculate the average accuracy and standard deviation for these 10 subjects. Due to space constraints, we are unable to present the detailed results for all 30 subjects in the CogPilot dataset. Instead, we report the average accuracy and standard deviation across all subjects. We observe a significant reduction in model accuracy compared to subject-dependent experimental setting, with the maximum difference reaching 21%. This finding is

not surprising, as physiological data vary greatly between individuals. As shown in the table, accuracy varies widely between subjects. Our approach achieves the best results in the majority of subjects. However, for certain individual subjects, our self-supervised method does not outperform the best fully supervised method, and the best fully supervised results are highlighted in bold. Compared with other self-supervised methods, our approach achieves the best results. Overall, our method has a statistically significant superiority, achieving the best results for the average between subjects in both datasets.

## VI. DISCUSSION

Utilizing multimodal physiological data, our objective is to devise a multimodal self-supervised model for detecting driving cognitive alertness. When the dataset contains a large number of samples with high-quality labels, fully supervised learning can achieve better results compared to self-supervised learning models. The collection process of physiological data is cumbersome and costly, and the labeling of samples is highly susceptible to subjective factors. Our model is constructed within a comprehensive self-supervised framework of mask modeling. The MSE evoked attention mechanism offers guidance on the underlying nature of the data, while the multimodal cascaded attention module enhances the fusion of multimodal physiological signals. Additionally, the multimodal uncertainty module improves the robustness of the model. We assessed the effectiveness of our model using the DriveDB and CogPilot datasets, employing two tuning methods. Our experimental results demonstrate that our model surpasses other methods in terms of performance. Ablation experiments further reveal that each module significantly contributes to the overall model’s efficacy.

In fully supervised experiments, our method achieved the best results on the DriveDB dataset, ranking second only to Inceptiontime [16] and DSN [18] on the CogPilot dataset. Importantly, self-supervised methods have demonstrated the ability to enhance model performance when compared to fully supervised approaches. On the DriveDB dataset, self-supervised methods such as TF-C [33], TS-TCC [28], and CPC [35], which are based on contrastive learning, yielded suboptimal results. On the CogPilot dataset, CPC [35] achieved better performance. The transformer mask modeling self-supervised frameworks, MultiMAE and TimeMAE, performed worse than self-supervised models based on contrastive learning. In general, the linear probing evaluation method demonstrated inferior performance compared to the fine-tuning evaluation method. This is because the linear probing evaluation method only trains the classification head. Ablation experiments demonstrated the significance of the multimodal uncertainty-aware module, which effectively mitigates individual differences between subjects by introducing randomness, thereby enhancing the model’s robustness. The remaining two components make equal contributions to the overall performance.

We also explored the effect of several key parameters on the model. The first aspect concerns the influence of the masking rate on model performance. Both datasets yielded

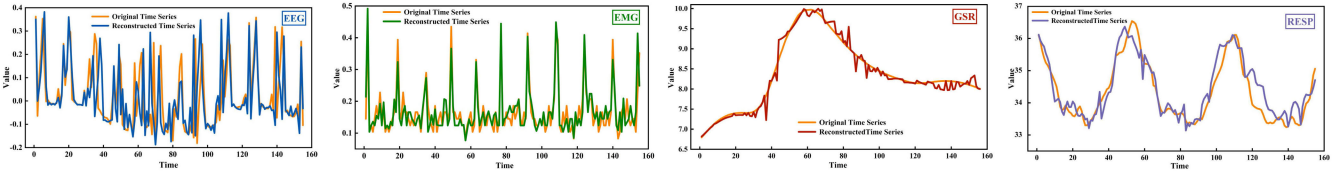


Fig. 8. Reconstructions of EEG, EMG, GSR and RESP from the DriveDB dataset.

TABLE IX  
COMPARISON AMONG OURS MODEL AND BASELINE METHODS IN SUBJECT-INDEPENDENT

Training strategy	Methods	DriverDB										CogPilot	
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Average	Average
Full-supervised	InceptionTime [16]	<b>75.83±1.12</b>	71.25±0.66	62.91±0.91	78.33±0.85	71.90±1.84	65.58±1.36	82.50±1.65	52.58±1.59	74.58±0.85	49.04±0.66	69.15±9.98	46.86±8.14
	ViT_Arjun [19]	74.16±1.07	60.00±1.15	65.83±1.34	77.91±1.22	72.85±1.69	55.83±2.34	76.66±0.91	51.66±0.75	76.72±1.86	62.38±1.22	67.40±9.56	44.44±10.6
	CrossMAE [20]	37.51±1.91	36.48±1.68	38.32±1.44	37.50±1.61	42.85±1.72	37.18±1.49	37.84±1.98	37.63±1.69	37.57±1.49	48.57±1.51	39.14±3.74	26.64±5.07
	Scinet [17]	70.83±0.79	66.66±0.92	<b>69.16±0.86</b>	78.33±0.56	70.47±0.61	60.41±0.73	75.41±1.37	53.83±0.64	74.58±1.28	58.09±1.02	68.47±6.89	49.76±9.28
	DSN [18]	67.91±0.50	74.58±0.42	65.41±0.84	80.25±0.30	72.38±0.71	58.33±0.25	80.41±0.35	53.33±0.95	76.25±0.88	<b>64.28±0.53</b>	69.31±9.11	54.76±10.4
	Ours (Random.Init)	65.08±1.53	71.51±1.30	64.16±1.12	80.00±0.59	73.80±0.77	60.19±0.32	79.58±0.41	50.00±0.26	77.50±0.73	63.09±1.79	68.99±9.50	55.71±8.93
Full-supervised: Fine-tuning	MultiMAE [14]	61.35±1.48	66.32±1.43	56.25±1.93	70.83±1.77	68.57±0.72	55.41±0.49	72.91±1.32	45.41±2.63	69.58±0.78	49.85±1.88	61.64±9.52	53.60±11.6
	TF-C [33]	67.06±1.44	70.58±1.95	63.33±1.72	79.44±1.41	71.29±1.22	59.61±0.55	73.75±1.61	52.19±0.29	77.91±1.79	63.35±1.43	67.85±8.46	56.03±9.37
	TimeMAE [25]	52.08±1.57	61.29±1.74	57.40±1.55	69.03±1.51	66.45±1.66	51.35±1.24	73.27±1.49	53.28±1.35	71.66±1.82	59.72±1.15	61.55±8.18	51.61±7.42
	TS-TCC [28]	60.74±1.72	71.48±1.94	61.44±1.45	78.39±1.53	69.64±1.48	60.83±2.38	78.39±1.58	50.19±1.59	76.25±1.75	59.25±1.28	66.69±9.57	56.52±9.97
	TS2Vec [34]	59.25±1.92	67.22±1.46	62.34±1.29	69.93±1.88	70.37±1.69	56.66±1.97	70.41±1.71	48.28±1.37	75.85±1.27	60.23±1.62	64.05±8.22	50.45±11.2
	CPC [35]	60.37±1.38	67.50±0.62	65.66±0.42	79.25±1.43	71.60±1.94	56.85±1.70	75.85±0.59	41.66±1.77	72.40±1.48	61.11±0.83	65.23±10.9	46.88±6.61
Ours	67.67±1.35	<b>75.18±1.29</b>	68.33±0.92	<b>80.95±0.85</b>	<b>74.28±0.72</b>	<b>66.96±1.54</b>	<b>82.91±0.49</b>	<b>54.37±0.31</b>	<b>78.33±1.26</b>	62.85±1.77	<b>71.18±8.82</b>	<b>57.98±9.28</b>	
Full-supervised: Linear probing	MultiMAE [14]	41.25±1.13	37.50±1.35	38.00±1.59	58.07±1.26	39.14±1.51	47.22±1.97	65.83±1.87	41.37±1.29	55.48±1.52	42.85±1.47	46.67±9.79	40.38±9.11
	TF-C [33]	<b>50.83±1.62</b>	46.85±1.67	44.44±1.75	67.93±1.81	43.18±1.42	48.43±1.33	61.11±1.58	41.29±1.76	54.45±1.27	44.62±1.59	50.31±8.56	43.70±8.77
	TimeMAE [25]	43.33±1.62	51.66±1.53	41.59±1.84	61.49±1.37	37.22±1.26	43.53±1.65	66.67±1.43	36.82±1.66	49.82±1.91	48.87±1.37	48.10±9.84	39.32±6.45
	TS-TCC [28]	46.29±1.87	51.15±1.44	45.92±1.34	69.44±1.12	42.17±1.57	49.16±1.38	66.25±1.92	41.25±1.56	57.75±1.79	<b>50.61±1.33</b>	51.99±9.62	42.54±12.2
	TS2Vec [34]	45.18±1.69	40.66±1.94	37.92±1.34	60.33±1.82	31.90±1.61	36.25±1.79	66.48±1.44	37.25±1.88	53.07±1.32	43.21±0.66	45.23±11.3	37.74±8.52
	CPC [35]	42.68±0.56	44.01±0.79	37.18±1.85	69.03±1.79	43.27±1.77	43.53±1.47	62.45±0.48	36.25±1.81	52.87±1.37	41.72±2.12	47.30±10.8	41.29±9.94
Ours	49.58±0.79	<b>52.91±1.59</b>	<b>46.66±1.64</b>	<b>69.58±1.76</b>	<b>45.71±0.62</b>	<b>49.92±1.37</b>	<b>67.08±1.47</b>	<b>42.91±0.85</b>	<b>58.71±1.59</b>	49.04±0.31	<b>53.16±9.02</b>	<b>45.21±9.27</b>	

optimal results at masking rates of 55% and 65%, respectively. A higher masking rate compels the model to enhance the encoder’s capabilities during the reconstruction process. It is important to note that the optimal masking rate cannot be determined empirically, as it varies depending on the dataset. Secondly, we also investigated the ratio of  $D_{train}$  during the fine-tuning phase. We found that using 40% of the  $D_{train}$  leads to better results. This demonstrated that our self-supervised pre-training model is well-suited for datasets with small sample sizes and sparse labels. Thirdly, we conducted experiments to assess the impact of encoders of different sizes on the model. While overall performance significantly improves as the model becomes larger and the number of layers increases, it’s essential to strike a balance between model performance and computational efficiency when selecting the optimal configuration. Finally, we explored the generalization performance of the model in a subject-independent setting. While our method achieved the best results for most patients, there remains significant room for improvement.

Even though the proposed study has introduced a novel framework that surpasses the performance of previous methods, it still exhibits specific limitations. Firstly, finding an optimal set of parameters for the model is a complex task as these parameters are interrelated. Secondly, each modality does not fully exploit its unique advantages, instead, there is a dynamic balance and interaction among them, which is achieved during the training process. Third, the model has not yet achieved optimal results in a subject-independent setting and requires further improvement and optimization.

VII. CONCLUSION

In this paper, we introduce a multimodal self-supervised model designed for the detection of driver cognitive alertness.

To delve into the intrinsic state of physiological signals, we innovatively introduce MSE and integrated it with an evoked attention. This approach serves to emphasize the effective information within the original data. In the process of multimodal fusion, we design a novel cascade attention mechanism. This mechanism not only accomplishes information interaction within each modality through self-attention but also sequentially integrates information across modalities. Importantly, the computational complexity of this attention is significantly reduced, being only 1/4 of the original self-attention. Furthermore, we introduce multimodal uncertainty sensing to effectively address the intricate variations in physiological signals within the target domain. We conduct extensive exploratory and comparative experiments on datasets related to both drivers and pilots. The experimental results unequivocally demonstrate that our model achieves superior performance, surpassing multiple self-supervised baselines. Our model has great potential for research and application in resource-limited scenarios. In our future endeavors, we aim to achieve two objectives: 1) Enhance the efficient fusion of multimodal physiological signals. 2) Develop an excellent pre-training model to improve the model’s performance in linear probing evaluation. 3) Improve the model to enhance its generalization ability across subjects.

REFERENCES

- [1] J. M. Mase, P. Chapman, and G. P. Figueredo, “A review of intelligent systems for driving risk assessment,” *IEEE Trans. Intell. Veh.*, pp. 1–17, Sep. 2024, doi: 10.1109/TIV.2023.3318113.
- [2] Z. Zhang, H. Ning, and F. Zhou, “A systematic survey of driving fatigue monitoring,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 19999–20020, Nov. 2022.
- [3] F. Zhou et al., “Driver fatigue transition prediction in highly automated driving using physiological features,” *Exp. Syst. Appl.*, vol. 147, Jun. 2020, Art. no. 113204.

- [4] B. Mandal, L. Li, G. S. Wang, and J. Lin, "Towards detection of bus driver fatigue based on robust visual analysis of eye state," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 545–557, Mar. 2017.
- [5] J. Fan, J. W. Wade, A. P. Key, Z. E. Warren, and N. Sarkar, "EEG-based affect and workload recognition in a virtual driving environment for ASD intervention," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 1, pp. 43–51, Jan. 2018.
- [6] J. R. Perello-March, C. G. Burns, S. A. Birrell, R. Woodman, and M. T. Elliott, "Physiological measures of risk perception in highly automated driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4811–4822, May 2022.
- [7] I. I. Esener, "A driver authentication system integrated to stress-level determination for driving safety," *Soft Comput.*, vol. 27, no. 15, pp. 10921–10940, Apr. 2023.
- [8] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.
- [9] S. Aouabdi, M. Taibi, S. Bouras, and N. Boutasseta, "Using multi-scale entropy and principal component analysis to monitor gears degradation via the motor current signature analysis," *Mech. Syst. Signal Process.*, vol. 90, pp. 298–316, Jun. 2017.
- [10] W. Li, X. Shen, Y. Li, and Z. Chen, "Improved multivariate multiscale sample entropy and its application in multi-channel data," *Chaos: Interdiscipl. J. Nonlinear Sci.*, vol. 33, no. 6, Jun. 2023, Art. no. 063125.
- [11] M. Bijelic et al., "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11679–11689.
- [12] H. Rivera-Flor, D. Gurve, A. Floriano, D. Delisle-Rodriguez, R. Mello, and T. Bastos-Filho, "CCA-based compressive sensing for SSVEP-based brain-computer interfaces to command a robotic wheelchair," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022.
- [13] J. Zhao, Y. Shi, W. Liu, T. Zhou, Z. Li, and X. Li, "A hybrid method fusing frequency recognition with attention detection to enhance an asynchronous brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2391–2398, 2023.
- [14] R. Bachmann, "Multimae: Multi-modal multi-task masked autoencoders," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 348–367.
- [15] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1578–1585.
- [16] H. I. Fawaz et al., "InceptionTime: Finding AlexNet for time series classification," *Data Mining Knowl. Discovery*, vol. 34, no. 6, pp. 1936–1962, Nov. 2020.
- [17] M. H. Liu et al., "SCINet: Time series modeling and forecasting with sample convolution and interaction," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 5816–5828.
- [18] Q. Xiao et al., "Dynamic sparse network for time series classification: Learning what to 'see,'" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 16849–16862.
- [19] A. Arjun, A. S. Rajpoot, and M. Raveendranatha Panicker, "Introducing attention mechanism for EEG signals: Emotion recognition with vision transformers," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 5723–5726.
- [20] Y. H. Zhang and J. C. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–21.
- [21] S. Z. Liu et al., "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–20.
- [22] A. D. Díaz Alonso, C. M. Travieso, J. B. Alonso, M. K. Dutta, and A. Singh, "Biometric personal identification system using biomedical sensors," in *Proc. 2nd Int. Conf. Commun. Control Intell. Syst. (CCIS)*, Nov. 2016, pp. 104–109.
- [23] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 2114–2124.
- [24] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," 2022, [arXiv:2211.14730](https://arxiv.org/abs/2211.14730).
- [25] M. Y. Cheng et al., "A time series is worth 64 words: Long-term forecasting with transformers," 2023, [arXiv:2303.00320](https://arxiv.org/abs/2303.00320).
- [26] Z. Li, Z. Rao, L. Pan, P. Wang, and Z. Xu, "Ti-MAE: Self-supervised masked time series autoencoders," 2023, [arXiv:2301.08871](https://arxiv.org/abs/2301.08871).
- [27] J. Dong, H. Wu, H. Zhang, L. Zhang, J. Wang, and M. Long, "SimMTM: A simple pre-training framework for masked time-series modeling," 2023, [arXiv:2302.00861](https://arxiv.org/abs/2302.00861).
- [28] E. Eldele et al., "Time-series representation learning via temporal and contextual contrasting," 2021, [arXiv:2106.14112](https://arxiv.org/abs/2106.14112).
- [29] O. B. Guney, M. Oblokulov, and H. Ozkan, "A deep neural network for SSVEP-based brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 2, pp. 932–944, Feb. 2022.
- [30] W. Wei, J. Zhou, H. Li, and Y. Wu, "ALUM: Adversarial data uncertainty modeling from latent model uncertainty compensation," 2023, [arXiv:2303.16866](https://arxiv.org/abs/2303.16866).
- [31] X. T. Li, Y. X. Dai, Y. X. Ge, J. Liu, Y. Shan, and L. Y. Duan, "Uncertainty modeling for out-of-distribution generalization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–13.
- [32] H. Rao et al., "Multimodal physiological monitoring during virtual reality piloting tasks (version 1.0.0)," *PhysioNet*, Aug. 2022, doi: [10.13026/azwa-ge48](https://doi.org/10.13026/azwa-ge48).
- [33] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 3988–4003.
- [34] Z. Yue et al., "TS2Vec: Towards universal representation of time series," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 8, pp. 8980–8987.
- [35] T. Mehari and N. Strodthoff, "Self-supervised representation learning from 12-lead ECG data," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105114.