Husformer: A Multi-Modal Transformer for Multi-Modal Human State Recognition

Ruiqi Wang[1] , Wonse Jo[1] , Dezhong Zhao[2], Weizheng Wang[1], Arjun Gupte[1], Baijian Yang[1], Guohua Chen[2], and Byung-Cheol Min[1]

***Abstract*—Human state recognition is a critical topic with pervasive and important applications in human-machine systems. Multi-modal fusion, which entails integrating metrics from various data sources, has proven to be a potent method for boosting recognition performance. Although recent multi-modal-based models have shown promising results, they often fall short in fully leveraging sophisticated fusion strategies essential for modeling adequate cross-modal dependencies in the fusion representation. Instead, they rely on costly and inconsistent feature crafting and alignment. To address this limitation, we propose an end-to-end multi-modal transformer framework for multi-modal human state recognition called *Husformer*. Specifically, we propose using cross-modal transformers, which inspire one modality to reinforce itself through directly attending to latent relevance revealed in other modalities, to fuse different modalities while ensuring sufficient awareness of the cross-modal interactions introduced. Subsequently, we utilize a self-attention transformer to further prioritize contextual information in the fusion representation. Extensive experiments on two human emotion corpora (DEAP and WESAD) and two cognitive load datasets (MOCAS and CogLoad) demonstrate that in the recognition of the human state, our *Husformer* outperforms both state-of-the-art multi-modal baselines and the use of a single modality by a large margin, especially when dealing with raw multi-modal features. We also conducted an ablation study to show the benefits of each component in *Husformer*. Experimental details and source code are available at: https://github.com/SMARTlab-Purdue/Husformer.**

***Index Terms*—Cognitive Load Recognition, Emotion Prediction, Multi-modal Fusion, Cross-modal Attention, Transformer**

I. Introduction

RECOGNITION of human states, such as affective states (commonly known as emotions) or cognitive load (often referred to as mental stress) plays a pivotal role in human-machine interaction systems [1], [2]. It enables machines to perceive, understand, and adapt to different human emotional or cognitive states, improving the performance of the whole systems [3], [4], [5]. The methods used to assess human emotions or cognitive load can be generally divided into two main categories based on the types of signals used: *physiological* and *behavioral* [6]. Physiological assessments involve measuring human physiological metrics, such as galvanic skin response (GSR), electroencephalography (EEG), electrooculography (EOG), electrocardiography (ECG), electromyogram (EMG), and heart rate (HR). These metrics change in response to involuntary reactions of the human nervous system under specific states [7]. On the other hand, behavioral assessments analyze subconscious human behavioral responses, including facial expressions, body and eye movements, and mouse movements, and associate them with different human states.

Unfortunately, due to the inherent complexity of human state reasoning, relying solely on signals from a single modality is unlikely to yield optimal recognition performance, especially in terms of accuracy and robustness [6], [8], [9]. Each modality exhibits different sensitivities to varying task environments and subject characteristics, rendering the identification of a universally effective modality for every scenario and individual unfeasible. Furthermore, a reliance on unimodal data sources increases susceptibility to noise and signal disruptions, which can lead to considerable inaccuracies or even complete system failures.

Recently, multi-modal fusion-based methods for human state recognition have emerged, combining data from various modalities to overcome the limitations of single-modal approaches [10], [11], [12], [13], [14]. The adoption of multi-modal signals can reduce the noise-to-signal ratio and enhance tolerance against sensor failures. More importantly, fusing different metrics collected from the same subject under one particular human state through multiple modalities can reveal important and comprehensive indexes of human emotion and cognitive load that are inaccessible via a single modality [6], [9]. Nevertheless, the inherent heterogeneity of multiple modalities poses challenges in generating an efficient fusion index of the human state. These challenges include: 1) the usual misalignment among different modalities, resulting in varied feature lengths and temporal resolutions; 2) the risk of incorporating biased or irrelevant information when merging multi-modal features due to noncommensurability across modalities; and 3) the necessity to infer long-term and complex dependencies across modalities for precise fusion [6].

Current multi-modal fusion approaches for human state recognition remain in their early stages and have not yet fully addressed the challenges arising from the heterogeneity across multiple modalities. Most methods rely on extensive feature engineering and alignment to concatenate features from different modalities and produce the fusion representation [6], [15], [16]. However, such direct concatenation fusion schemes often overlook the latent correlations across modalities and may still face limitations due to non-instantaneous coupling, even when manually aligned. While there are methods to facilitate learning of cross-modal interactions through shared representations [14], [17], [18], [19], [20], these approaches

[1]Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA [wang5357, jow, wang5716, guptea, byang, minb] @purdue.edu

[2]College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, Beijing, China. DZ_Zhao@buct.edu.cn, chengh@mail.buct.edu.cn

† Equal contribution

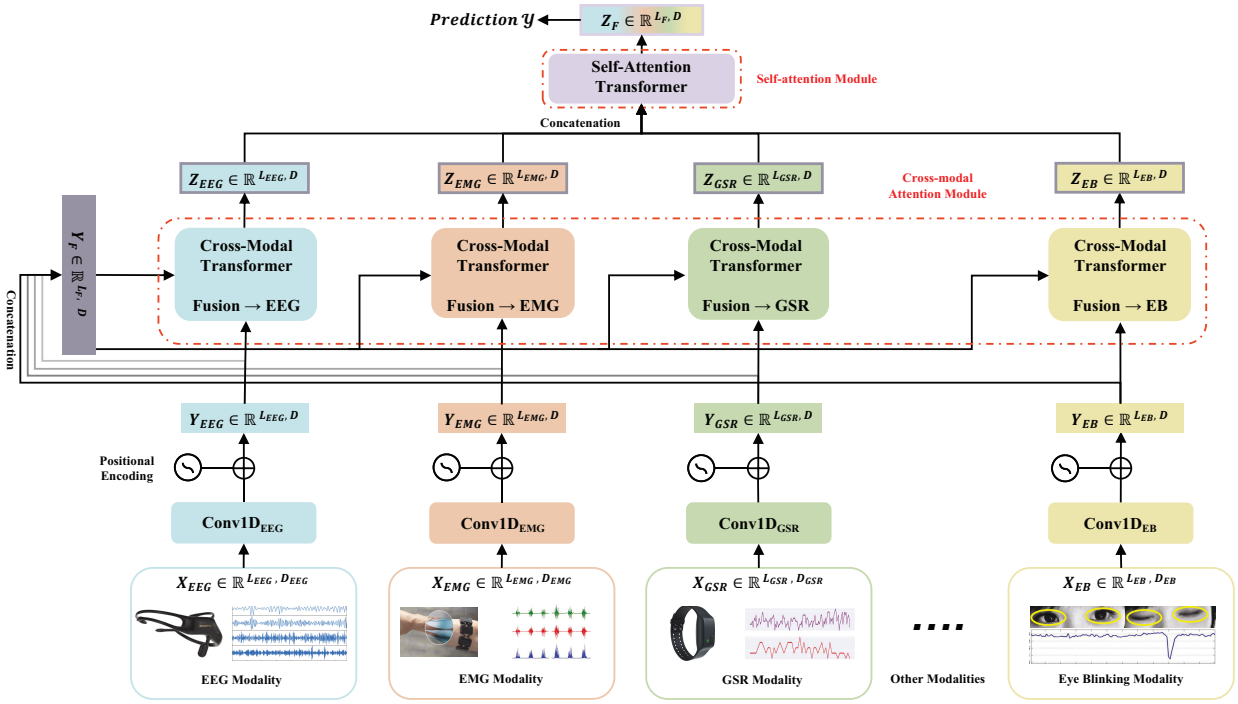


Fig. 1. Framework of the proposed *Husformer*, taking multi-modal data input from four example modalities: EEG, EMG, GSR, and eye blinking (EB). The multi-modal data inputs, $X_{(.)} \in \mathbb{R}^{L_{(.)}, D_{(.)}}$, where $L_{(.)}$ and $D_{(.)}$ separately present the length and dimension of the input sequence of one modality, are passed through multiple one-dimension temporal convolution layers, $Conv1D_{(.)}$ (Section III-B), and then encoded with positional information to produce the low-level unimodal features $Y_{(.)} \in \mathbb{R}^{L_{(.)}, D}$ which all have the same dimension D (Section III-C); these are concatenated to generate the low-level fusion representation $Y_F \in \mathbb{R}^{L_F, D}$. This representation is then fed alongside the unimodal low-level features of each modality into each respective cross-modal attention transformer, wherein the target modality is adapted and reinforced according to the other resource modalities through learning the attention between its unimodal features and the low-level representation (Section III-D). Then all reinforced unimodal features $Z_{(.)} \in \mathbb{R}^{L_{(.)}, D}$ are concatenated into the mid-level fusion representation, which is passed through a self-attention transformer to generate the high-level fusion representation $Z_F \in \mathbb{R}^{L_F, D}$ with important contextual information prioritized. Finally, the high-level fusion representation is transported to fully connected layers to make predictions (Sections III-E and III-F).

typically employ network structures that are insufficiently deep to effectively grasp complex cross-modal dependencies and establish necessary complementarity among different modalities. Moreover, the prerequisite feature engineering processes in these methods are not only resource-intensive but also subject to variability across different modalities. Parameter optimization in such contexts requires significant expert knowledge and thorough cross-validation [7], [21], which undermines the simplicity of the model and its applicability to new task scenarios. Additionally, to the best of our knowledge, existing multi-modal methods have not been demonstrably effective in predicting both human affective states and cognitive loads. Most are designed for specific modality combinations and task scenarios, which raises concerns about their general applicability.

To address the aforementioned gaps, we introduce *Husformer*: an end-to-end multi-modal trans*former* designed for *hu*man *s*tate recognition. The model can efficiently learn representations of human emotion or cognitive load from heterogeneous multi-modal data streams. Fig. 1 illustrates its structure with four example input modalities: EEG, EMG, GSR, and eye blinking (EB). The core components of *Husformer* are the cross-modal attention module and self-attention module, which consist of multiple cross-modal attention transformers and one self-attention transformer, respectively. The cross-modal attention transformers model the latent interactions across modalities by continuously adapting and reinforcing features from one modality with those of other modalities (e.g., EEG ← EMG, GSR, and EB). Unlike direct concatenation of multi-modal modalities or learning cross-modal shared representations through shallow neural networks, our cross-modal attention mechanism encourages the target modality to directly attend to low-level features in other modalities, where strongly relevant and complementary information is revealed. This leads to more adaptive and efficient complementarity and cooperation across multiple modalities. The subsequent self-attention transformer prioritizes important contextual information in the fusion representation concatenated from reinforced unimodal features of all modalities supplied by the cross-modal attention transformers. This finally generates a weighted high-level fusion representation based on which predictions are made.

To evaluate the performance of *Husformer*, we conducted extensive experiments on four multi-modal datasets: DEAP [22] and WESAD [23] for emotion recognition, and MOCAS [24] and CogLoad [25] for cognitive load estimation. Additionally, we performed a comprehensive ablation study to investigate the benefits of each module in *Husformer*.

The main contributions of this work can be summarized as follows:

- *Husformer* is an end-to-end model that learns directly and efficiently from heterogeneous multi-modal physiological

and behavioral signals without the massive feature crafting and alignment required in previous works.

We introduce cross-modal attention transformers to fuse features from different modalities and sufficiently model long-term cross-modal interactions, and then a self-attention transformer to prioritize effectual contextual information in the fusion representation.

To the best of our knowledge, this is the first time a generic model for human state recognition is presented and proven to be effective for both human emotion and cognitive load. Our extensive experiments on four publicly available datasets demonstrate the benefits of the *Husformer* and each of its constituent modules.

II. Background

This section reviews existing research on human state recognition using multi-modal fusion approaches and the preliminary transformer networks that serve as the basis of our model.

A. Multi-Modal Fusion for Human State Recognition

The concept of 'human state' refers to either the emotional state or cognitive load borne by an individual. Given the complex nature of the human state, a single modality is insufficient to achieve recognition with satisfactory performance in terms of accuracy and robustness, especially in real-world task scenarios where signals are more subject to interruption, noise, and delay [9]. To solve this issue, the multi-modal fusion that integrates human signals from more than one source modality into a synchronized compact representation has been adopted.

Nevertheless, multi-modal fusion methods for human state recognition are still at the initiatory stages and suffer from several defects in fusion strategies. Firstly, the direct concatenation of different modalities at the sensor, feature, or decision level has been adopted for most existing works to produce the fusion representation [6], [15]. This approach may introduce superfluous information, bias, or noise due to feature noncommensurability across different modalities. For example, the mere fusion of heart rate (HR) with other modalities might offer additional valuable information in some tasks but cause disruption in others. This is because HR-related signals are influenced not only by cognitive load and emotion, but also by irrelevant physical activities [6]. Moreover, the direct fusion of different modalities with unbalanced length and temporal dimensions necessitates extensive feature alignment preprocessing. This could diminish the richness and diversity of the input data, potentially constraining the model's capacity to discern complex patterns and relationships within or across modalities. In general, simple concatenation operations are insufficient for modeling cross-modal interactions, likely failing to uncover crucial representations that multi-modal fusion could otherwise reveal.

Recently, advanced fusion schemes have emerged, aiming to address these challenges by modeling cross-modal interactions through the training of shared representations across different modalities. For instance, Tang *et al.* [14] employed Restricted Boltzmann Machines (RBM) [26] to train a hidden layer, anticipated to learn the shared representations of sub-layers from diverse modalities. Tsai *et al.* [27] introduced cross-modal temporal correlations by learning shared weights across sub-layers of different modalities. Qiu *et al.* [18] leveraged Deep Canonical Correlation Analysis (DCCA) [28] to amplify the correlations among features of two modalities. Nonetheless, these methods, utilizing shallow network structures, may not thoroughly capture intricate correlations across diverse modalities, particularly from raw multi-modal signals where features are more disjointed. Additionally, the DCCA used by Qiu *et al.* [18] and Liu *et al.* [19] is confined to analyzing the correlation between only two modalities, constraining the model's applicability to general task scenarios that may necessitate the concurrent fusion of more than two modalities

Furthermore, existing works usually require extensive feature crafting and alignment procedures to reach sound recognition performance due to the abundant collinearity or heterogeneity in the raw multi-modal features. In addition, different modalities are usually preprocessed by different methods whose optimal parameters are initially unknown. For instance, Zhou *et al.* [29] employed infinite impulse response (IIR) high-pass and Hanning window filters for EEG modality feature engineering, while GSR and HR modalities were processed using continuous decomposition analysis and other physiological methods. Subsequently, procedures like independent component analysis (ICA) and manual feature selection based on prior knowledge were utilized for a second round of preprocessing. Such intricate, non-uniform feature engineering procedures detract from the model's simplicity and universality, rendering it less suitable for general task scenarios and particularly for real-world applications. This contradicts the fundamental aim of multi-modal fusion. Moreover, to our knowledge, existing studies primarily focus on recognizing either affective or cognitive states, often tailored to specific tasks and not validated on publicly available datasets. Consequently, these models lack the generality and replicability essential for broader applicability.

Distinct from these prior studies, our proposed *Husformer* 1) focuses on the general recognition of human emotion or cognitive load; 2) does not require extensive feature alignment and extensive feature crafting procedures in existing works, but rather learns from raw multi-modal feature streams; and 3) utilizes cross-modal attention layers as the fusion strategy, thereby introducing efficient cooperation and complementary adaptions across modalities regardless of the number of modalities.

B. Transformer Network

The transformer network [30] was originally proposed to solve sequence-to-sequence machine translation tasks in the natural language processing area. Unlike the traditional encoder-to-decoder structure, the transformer network adopts a multi-head self-attention mechanism to substitute for the attention-based convolution and recurrence layers. The self-attention mechanism aims to calculate a global representation of a sequence input that reveals meaningful contextual information by relating different components within the sequence.

A self-attention block adapts each entity in a sequence by considering the global contextual information of the whole sequence. Multi-head self-attention splits the attention into multiple latent sub-spaces (heads), which enables the modeling of multiple complex contextual relations across elements in the sequence, leading to a more comprehensive global representation.

III. Approach

In this section, we present *Husformer*, an end-to-end multi-modal transformer for multi-modal human affective or cognitive state recognition that learns the fusion representation directly and efficiently from multi-modal data streams.

A. Overview

As shown in Fig. 1, at a high level, the *Husformer* employs a position-wise feed-forward process to merge multi-modal signal series from multiple cross-modal transformers (Section III-D). Within each cross-modal transformer, the target modality is continuously enriched with low-level features from other modalities by computing the latent cross-modal attention between the target low-level unimodal feature and the low-level fusion representation. A sequence-to-sequence model, specifically a self-attention transformer, is then used to handle the mid-level fusion representation sequence, which includes all enhanced unimodal features, producing the adaptively weighted high-level fusion representation (Section III-E). This is achieved by computing multi-head self-attention (Section II-B), where the self-attention transformer examines the pairwise relationships across elements in the mid-level fusion representation, namely the bolstered unimodal features, to calculate adaptive weights at various positions, thereby emphasizing essential contextual information. Ultimately, the high-level fusion representation undergoes processing through fully connected layers for prediction (Section III-F).

B. Temporal Convolutions

Let , , denote modalities. And let present the raw multi-modal data sequences input from these modalities. and represent the sequence length (e.g., channel number) and dimension (e.g., sampling rate) of each unimodal input respectively in this paper. The multi-modal input sequences are passed through multiple one-dimension temporal convolution layers with different kernels to generate multiple convoluted sequences with the same dimension :

(1)

where denotes the temporal convolution kernel sizes for modalities , , .

Each convoluted sequence aims to contain low-level temporal features of each modality. Furthermore, it is important that after temporal convolutions, different unimodal input sequences are projected to the same dimension, making the dot-product calculation in the following cross-modal attention module mathematically feasible.

C. Positional Encoding

As mentioned in the introduction of the transformer network (Section II-B), the transformer model has no inherent awareness of the positional information of each sequence component, such as the relative or absolute position of features within a modality sequence. To introduce sufficient awareness of relations across neighboring elements, i.e., features of adjacent channels within one modality sequence, and thus spatial information, we follow the method proposed in [30] and apply positional encoding (PE) to the convoluted sequences using and functions with different frequencies. The PE of one convoluted sequence can be defined as a matrix:

(2)

where and — .

Each characteristic dimension (i.e., column) of is a position index displayed in the sinusoidal pattern. The calculated PEs , are then augmented with convoluted sequences to obtain low-level unimodal feature sequences with both initial temporal and spatial information encoded :

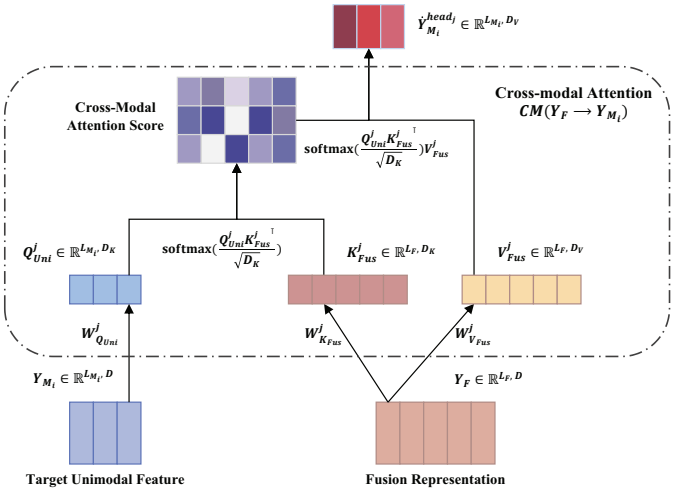
(3)

Additionally, the extracted low-level unimodal feature sequences of all modalities are then concatenated to produce the low-level fusion representation :

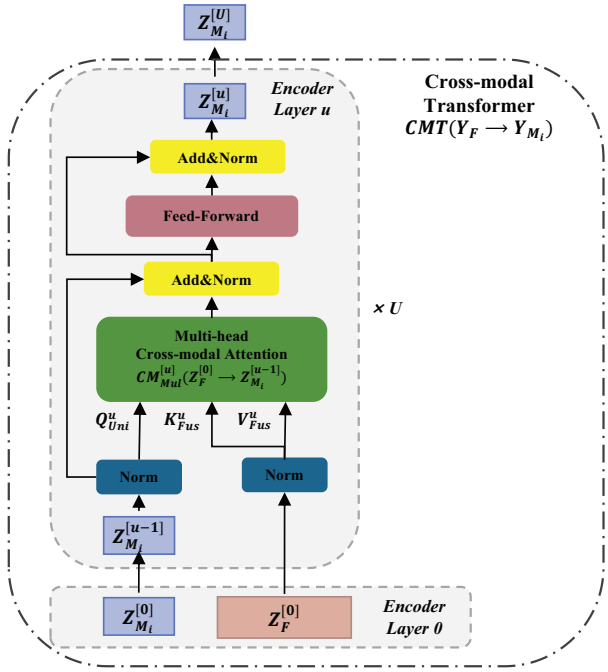
(4)

D. Cross-modal Attention Module

To provide sufficient complementary interactions and adaptions across different modalities, we respectively feed the low-level unimodal feature sequence of each modality with the low-level multi-modal fusion representation to a cross-modal attention module that is comprised of multiple cross-modal transformer networks. Each cross-modal transformer is designed to continuously bolster the low-level unimodal features of the target modality using features from other source modalities. This is achieved by learning cross-modal attention between the input unimodal sequence and the low-level fusion representation. Essentially, this learned attention encourages each target modality to directly access the low-level features of other source modalities encoded in the fusion representation. This process adaptively identifies relevant and beneficial information that can act as complementary reinforcements. This module not only promotes a nuanced understanding of correlations across modalities, but also mitigates potential artifacts and feature inconsistencies in raw multi-modal features. It achieves this by selectively ignoring irrelevant components, such as disrupted or non-responsive features in the low-level unimodal feature sequence, during cross-modal attention computation.



(a) Illustration of the cross-modal attention between each low-level unimodal feature and the low-level fusion representation.



(b) Illustration of the cross-modal Transformer network stacked by multiple identical cross-modal attention encoder layers.

Fig. 2. Architectural description of cross-modal attention and cross-modal transformer network.

1) Cross-modal Attention: The purpose of our cross-modal attention is to learn an attention score between a target low-level unimodal feature sequence $Y_{M_i} \in \mathbb{R}^{L_{M_i}, D}$ and the low-level multi-modal fusion representation $Y_F \in \mathbb{R}^{L_F, D}$, which guides the adaption and reinforcement for the target unimodal features using other source unimodal features embedded in the fusion representation. We formulate unimodal Query Q_{Uni}, fusion Key K_{Fus} and Value V_{Fus} as:

$$\begin{aligned} Q_{Uni} &= Y_{M_i} \cdot W_{Q^{Uni}} \\ K_{Fus} &= Y_F \cdot W_{K^{Fus}} \\ V_{Fus} &= Y_F \cdot W_{V^{Fus}} \end{aligned} \tag{5}$$

where $W_{Q^{Uni}} \in \mathbb{R}^{D, D_Q}$, $W_{K^{Fus}} \in \mathbb{R}^{D, D_K}$ and $W_{V^{Fus}} \in \mathbb{R}^{D, D_V}$ are learnable weights.

As depicted in Fig. 2a, similar to the self-attention process described in [30], the latent adaption and reinforcement from the fusion representation to the target unimodal feature, i.e., the learned cross-modal attention $\dot{Y}^{head_j}_{M_i} \in \mathbb{R}^{L_{M_i}, D_V}$, in the j^{th} head cross-modal attention can be defined as:

$$\begin{aligned} \dot{Y}^{head_j}_{M_i} &= CM(Y_F \rightarrow Y_{M_i}) \\ &= Attention(Q^j_{Uni}, K^j_{Fus}, V^j_{Fus}) \\ &= \text{softmax}\left(\frac{Q^j_{Uni} \cdot {K^j_{Fus}}^\top}{\sqrt{D_K}}\right) \cdot V^j_{Fus} \end{aligned} \tag{6}$$

where the $\text{softmax}(.) \in \mathbb{R}^{L_{M_i}, L_F}$ presents the scaled cross-modal attention score matrix between the fusion representation and the target unimodal feature.

We define the $\dot{Y}^{head_j}_{M_i}$ in (6) as the single-head cross-modal attention. Accordingly, the multi-head cross-modal attention between the i^{th} target modality and the fusion representation can be formulated as:

$$\begin{aligned} \dot{Y}^{Mul}_{M_i} &= CM_{Mul}(Y_F \rightarrow Y_{M_i}) \\ &= \text{Concat}\left(\dot{Y}^{head_1}_{M_i}, \ldots, \dot{Y}^{head_m}_{M_i}\right) \end{aligned} \tag{7}$$

where $\dot{Y}^{Mul}_{M_i} \in \mathbb{R}^{L_{M_i}, mD_V}$ and m is the head number.

2) Cross-modal Transformer: Drawing upon the structure of the self-attention transformer network as detailed in [30], we have developed the cross-modal transformer by integrating the previously defined multi-head cross-modal attention. As illustrated in Fig. 2b, the cross-modal transformer, denoted as $CMT_{Y_F \rightarrow Y_{M_i}}$, processes the i^{th} target unimodal feature sequence alongside the low-level fusion representation. It consists of several identical layers, each comprising a multi-head cross-modal attention block paired with a position-wise feed-forward network. It also includes residual connections and layer normalization for enhanced efficiency and stability. Formally, a cross-modal transformer network with U cross-modal attention encoder layers calculates the reinforced unimodal feature sequence $Z_{M_i} \in \mathbb{R}^{L_{M_i}, D}$ feed-forwardly as:

$$\begin{aligned} Z^{[0]}_F &= Y_F \\ Z^{[0]}_{M_i} &= Y_{M_i} \\ \hat{Z}^{[u]}_{M_i} &= CM^{[u]}_{Mul}(\mathbb{L}(Z^{[0]}_F) \rightarrow \mathbb{L}(Z^{[u-1]}_{M_i})) \\ \dot{Z}^{[u]}_{M_i} &= \hat{Z}^{[u]}_{M_i} + \mathbb{L}(Z^{[u-1]}_{M_i}) \\ Z^{[u]}_{M_i} &= \mathbb{F}_\theta(\mathbb{L}(\dot{Z}^{[u]}_{M_i})) + \mathbb{L}(\dot{Z}^{[u]}_{M_i}) \end{aligned} \tag{8}$$

where $u \in [1, U]$ presents the u^{th} cross-modal attention encoder layer, $\mathbb{L}(.)$ and $\mathbb{F}_\theta(.)$ denote the layer normalization operation and position-wise feed-forward network with a parameter set θ, respectively.

In each cross-modal transformer, the target unimodal features are continuously encoded and enhanced with external information from other source unimodal features embedded in the fusion representation. Specifically, the low-level unimodal features of source modalities from the low-level fusion representation are converted to different pairs of the fusion Keys and Values in (5) to compute the multi-head cross-modal attention of the target modality in (7). Following the process

described in (8), each target modality is merged with other source modalities by a position-wise feed-forward process from each cross-modal transformer.

E. Self-attention Module

The outputs of the cross-modal attention module, namely the enhanced unimodal features for each modality, are aggregated into a single sequence. This forms the mid-level fusion representation , which is then fed into the self-attention transformer [30]. The role of the self-attention transformer here is to create a weighted high-level fusion representation , wherein essential contextual information is accentuated.

Specifically, the self-attention transformer dynamically assigns self-attention scores across different positions within the input sequence. This process involves each minimal unit of the enhanced unimodal features of various modalities, incorporating the global contextual information of the entire sequence. Consequently, the elements within the sequence are endowed with adaptive weights, thereby constructing a high-level global representation. In this representation, critical contextual information pertinent to human state recognition, such as sensitive and efficient features in specific task scenarios, is underscored, while less relevant features, like those disrupted and insensitive, are de-emphasized. This adaptive self-attention mechanism effectively mitigates potential feature noncommensurability across modalities, thereby augmenting the efficiency of the model.

F. Model Training

At the final step, the output of the self-attention module, namely, the high-level global feature , is passed through two linear layers with a residual connection operation and a softmax nonlinear activation function that calculates prediction probabilities as:

(9)

where and present two linear layers with parameter sets and , and contains the prediction probability for each class.

To reduce the class imbalance resulting from biased data distribution and varied recognition difficulty, we adopt the multi-class focal loss function [31] for training, which is formulated as:

(10)

where is the total number of classes, and correspond to the true label and predicted probability for class respectively, and present the balancing parameter that controls the trade-off between the positive and negative samples within class and a focusing parameter that down-weights the contribution of well-classified samples respectively.

The overall procedures of *Husformer* model training are summarized in Algorithm 1.

Algorithm 1 Procedures of *Husformer* Training

Given multi-modal data series of modalities and true classification labels
Given training steps
Initialize convolution kernel sizes , cross-modal attention weights , self-attention weights and other model parameters

// Convolutions and Positional Encoding
Compute low-level unimodal feature sequences by (1)-(3)
Compute low-level fusion representation by (4)
while **do**
    // Cross-modal Attention Module
    Compute reinforced unimodal feature sequences by (8)
    // Self-attention Module
    Compute high-level fusion representation
    // Linear Layers
    Compute prediction probability by (9)
    // Model Optimization
    Compute loss by (10)
    Update model parameters using backpropagation

end while

IV. Experimental Setting

In this section, we detail the experiments conducted on four publicly available multi-modal datasets, which are widely recognized in the field of human affective and cognitive state recognition. Our experiments are designed to benchmark the performance of our *Husformer* against five state-of-the-art baselines in multi-modal fusion-based human state recognition. To comprehensively assess the efficacy of our approach, we also report the performance achieved using single modalities on each dataset. This comparison aims to ascertain whether our multi-modal fusion-based *Husformer* surpasses the recognition capabilities of methods relying on individual modalities. Furthermore, an ablation study was conducted to elucidate the individual contributions and advantages of each module within the *Husformer* framework.

A. Datasets

We selected two multi-modal affective datasets: DEAP [22] and WESAD [23], and two multi-modal cognitive load datasets: MOCAS [32] and CogLoad [25] for experiments. For each dataset, we used the original features of each modality as provided, without any further feature alignment or engineering.

The DEAP dataset contains multiple physiological signals, including EEG, EMG, EOG, and GSR, collected from 32 participants watching 40 different music video clips that elicited different emotional states. After each video clip, participants were requested to report their affective state levels in terms of arousal, valence, liking, and dominance from 1 to 9 using the Self-Assessment Manikin (SAM). In our

TABLE I
DESCRIPTION OF UTILIZED MODALITIES IN THE DEAP DATASETS. FREQUENCY: TIME SAMPLING RATE; CHANNELS: NUMBER OF CHANNELS; AND ARRAY SHAPE: NUMBER OF DATA ROWS×CHANNELS×FREQUENCY.

Raw DEAP Dataset			
Modality	Frequency	Channels	Array Shape
EEG	512	32	69535×32×512
EMG	512	4	69535×4×512
EOG	512	4	69535×4×512
GSR	512	1	69535×1×512
Preprocessed DEAP Dataset			
Modality	Frequency	Channels	Array Shape
EEG	128	32	80640×32×128
EMG	128	2	80640×2×128
EOG	128	2	80640×2×128
GSR	128	1	80640×1×128

TABLE II
DESCRIPTION OF UTILIZED MODALITIES IN THE WESAD DATASET

WESAD Dataset			
Modality	Frequency	Channels	Array Shape
GSR (chest)	700	1	27287×1×700
BVP (wrist)	64	1	27287×1×64
EMG (chest)	700	1	27287×1×700
ECG (chest)	700	1	27287×1×700
RESP (chest)	700	1	27287×1×700
GSR (wrist)	4	1	27287×1×4

experiment, we utilized two versions of the DEAP dataset: the downloaded *Data-original.zip*, which contained collected raw multi-modal features, was regarded as the raw DEAP dataset, while the downloaded *Data-preprocessed.zip*, which crafted features with several procedures[1], was regarded as the preprocessed DEAP dataset. Valence and arousal were selected as the evaluation criteria of human emotion, where we mapped the scales (1-9) into three levels: "negative" or "passive"(1-3); "neutral" (4-6); and "positive" or "active" (7-9).

The WESAD dataset contains physiological data, consisting of GSR, BVP, EMG, ECG and respiration (RESP), collected by one chest-worn and one wrist-worn wearable sensor from 15 participants who conducted different tasks that aimed to elicit different emotional states. Specifically, participants were asked to close their eyes for seven minutes to stimulate the neutral state. The stress state was elicited by the Trier Social Stress Test (TSST) [33], where participants delivered a five-minute speech on their personal traits to three-person panels. Moreover, for the amusement state, participants were required to watch funny videos for 392 seconds. After each task, participants reported subjective emotional states using SAM and other self-report questionnaires, and three kinds of labels, i.e., neutral *vs.* stress *vs.* amusement, were provided.

The MOCAS dataset contains physiological data, including 5-channel EEGs, EEG band powers (or EEG_POW, including theta, low and high beta, alpha, and gamma bands of 5-channel EEGs), BVP, GSR and HR, and behavioral data, including Eye Aspect Ratio (EAR) and Action units (AUs), from 21 participants conducting Closed-Circuit Television (CCTV) monitoring tasks that aimed to elicit different levels of cognitive load. After each task, participants were required to report subjective cognitive load via NASA-TLX. Based on the weighted NASA-TLX scores, three categories of annotations, i.e., low *vs.* medium *vs.* high, cognitive load were given. Apart from the raw multi-modal physiological and behavioral features collected from two off-the-shelf wearable sensors: Empatica E4 and Emotiv Insight, and a webcam, the MOCAS also contains the data prepossessed by NeuroKit2 [34] and other methods [32]. In this experiment, we utilized both raw and preprocessed MOCAS datasets.

[1]https://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html

TABLE III
DESCRIPTION OF UTILIZED MODALITIES IN THE MOCAS DATASETS

Raw MOCAS Dataset			
Modality	Frequency	Channels	Array Shape
EEG	128	6	215341×5×128
EEG_POW	8	25	215341×25×8
BVP	128	1	215341×1×128
GSR	6	1	215341×1×6
EAR	1	1	215341×1×1
Preprocessed MOCAS Dataset			
Modality	Frequency	Channels	Array Shape
EEG	128	6	215341×5×128
EEG_POW	8	25	215341×25×8
BVP	128	1	215341×1×128
GSR	6	1	215341×1×6
EAR	1	1	215341×1×1

TABLE IV
DESCRIPTION OF UTILIZED MODALITIES IN THE COGLOAD DATASET

CogLoad Dataset			
Modality	Frequency	Channels	Array Shape
HR	1	1	89225×1×1
IBI	1	1	89225×1×1
GSR	1	1	89225×1×1
SKT	1	1	89225×1×1
ACC	1	2	89225×2×1

The CogLoad dataset contains physiological signals, including HR, IBI, GSR, SKT, and motion data (ACC) collected from 23 participants through a Microsoft band. The participants conducted six dual tasks including primary and secondary cognitive-load tasks that were expected to stimulate target levels of cognitive load. The primary task was randomly selected from six psycho-physiological tests proposed in [35]. The secondary task was to click on the appearing target on the screen while conducting the primary task. After each task, participants were asked to report subjective cognitive load based on the TLX mental demand of the NASA-TLX questionnaire, and the data collected in the baseline (rest) section was labeled as -1. In our experiment, we mapped the subjective scales into three classes of labels: low (-1-3), medium (4-6), and high (7-9).

In our experiments, to simulate realistic application scenarios, the input sample of each modality is a 2-D feature matrix extracted from a 1-second segment with the dimension of , i.e., the channel number plus the sampling frequency of the modality. The details of the modalities used in the aforementioned datasets are described in Tables I-IV.

B. Baselines

We selected five state-of-the-art baselines of multi-modal fusion-based human state recognition as the comparisons with our proposed *Husformer*:

EF-SVM: Support Vector Machines (SVMs) with early fusion [36], [37], [21]. This is a popular benchmark model

for human state recognition, which concatenates different modalities together at the sensor or feature level and builds an SVM as the classifier for the fused representation.

LF-SVM: SVMs with late fusion [38], [37], [21]. This model is also a strong benchmark to predict the human state, which fuses different modalities at the decision level. Each modality is processed with an SVM classifier to make individual predictions, which are combined together through a voting scheme. In our experiment, we select Dempster-Shafer Theory (DST) voting [39], [40], which is reported as the best-performing voting scheme by [21], as the late fusion process.

EmotionMeter [17]. This is a multi-modal deep learning model for human emotion prediction. Multiple individual RBMs are built to process features of each modality, where the hidden layers of those individual RBMs are concatenated together to learn the shared representations across different modalities. Then a linear SVM is adopted for classification using the learned shared representations.

MMResLSTM: Multimodal Residual Long Short-Term Memory Neural Network [41]. This is another state-of-the-art multi-modal deep learning model for human emotion recognition. Multiple individual LSTM blocks are constructed for each modality, where each LSTM layer of these individual LSTM blocks shares the same weights to learn the temporal correlations across different modalities. Finally, the outputs of all LSTM blocks are concatenated and passed through a dense layer to make predictions. Layer normalization and residual connection are also applied to accelerate the training process.

MMFN: Multi-modal fusion network with complementarity and importance for emotion recognition [20]. This is a recent deep learning model for multimodal human emotion recognition, utilizing the dot-product attention mechanism to discern both the complementarity and importance across different modalities. Within the *MMFN*, features from each modality are initially introduced into an *importance attention module*. In this module, each set of unimodal features is multiplied by a learnable corresponding importance weight, creating multiple unimodal representations. These representations are subsequently forwarded to a *complementary attention module*. This stage begins by forming feature matrices that encapsulate the relationship between each two modalities. The module then determines complementary weights by applying the softmax function to these matrices and conducting element-wise multiplication with the original feature vectors of each modality. These complementary vectors from each modality are merged and supplied to a bidirectional LSTM network for emotion prediction.

Moreover, to comprehensively evaluate if our *Husformer* could improve the performance compared with using single modality, we also implemented Long Short-Term Memory Nerual Network (LSTM) [42], Graph Neural Network (GNN) [43] and Transformer [30], which were broadly utilized for single-modal-based human state recognition [44], [45], [46], to test every single modality in each dataset, and the best-performing results among these three models were reported.

C. Evaluation and Metrics

For each dataset, we randomly shuffled all data and conducted the -folder cross-validation ( ) [47], [48]. We reported the average multi-class accuracy ( ) and multi-class average F1-score ( ) [49] with standard deviations for each model in each experiment. Furthermore, samples from the same trials, such as a film clip in the DEAP dataset or a monitoring task in the MOCAS dataset, were included in either the training or test sets. We also conducted two-sample independent t-tests to compare the classification results of different models, and the significance is asserted when .

D. Implementation Details

All training and experiments were conducted on an NVIDIA Tesla V100 GPU. We trained all baseline networks by following the implementation procedures described in their respective original papers. Also, for the SVM classifier, we followed the approach described in [21] to select the Radial Basis Function (RBF) kernel and optimize the values of C and . Note that *EF-SVM* cannot be directly applied to unaligned datasets, which means datasets that contain multiple modalities with different time sampling rates, since the concatenation operation is mathematically impossible due to different feature dimensions. Therefore, we added multiple one-dimensional convolution sub-networks with the same structures and parameters as those in the *Husformer* before the *EF-SVM* to extract low-level unimodal features with the same dimension for the concatenation operation on unaligned datasets, i.e., WESAD and MOCAS datasets. The detailed information regarding the selection and settings of the *Husformer* hyperparameters utilized in our experiments can be found in Appendix A.

E. Ablation Study

To evaluate the benefits of each module in our *Husformer*, we also built three ablation models for the ablation study:

HusFuse: Deleting the cross-modal attention module in the *Husformer*, and fusing the low-level features of all modalities directly to the self-attention module.

HusLSTM: Replacing the self-attention transformer in the *Husformer* with an LSTM layer, which is widely used as a backbone network in emotion or cognitive load prediction tasks.

HusPair: Replacing the cross-modal attention module in the *Husformer* with the directional pairwise cross-modal attention widely adopted in multi-modal natural language processing and computer vision areas [50], [27]. However, *Husformer* differs from *HusPair* in that it focuses on the cross-modal attention between each individual modality and the multi-modal fusion signal, rather than between one single modality and another. This allows the model to consider the coordination of more than a pair of modalities at the same time, reducing the potential information redundancy caused by parallel pairwise fusion.

Furthermore, to ensure a fair comparison, we kept the hyper-parameters of ablation models the same as those in the *Husformer* during the experiments.

TABLE V
PERFORMANCE OF DIFFERENT MODELS ON RAW DEAP AND PREPOSSESSED DEAP DATASETS IN TERMS OF AVERAGE MULTI-CLASS AVERAGE ACCURACY (Acc) AND MULTI-CLASS AVERAGE F1-SCORE ($F1$) WITH STAND DEVIATIONS. RESULTS OF OTHER MODELS THAT ARE WITHIN 5% OF *Husformer*'S PERFORMANCE ON Acc OR $F1$ ARE HIGHLIGHTED. ^{h}: HIGHER VALUES INDICATE BETTER PERFORMANCE.

Dataset	Raw DEAP				Preprocessed DEAP			
Criteria	Valence		Arousal		Valence		Arousal	
Metric	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$
EF-SVM	43.95±2.17	47.36±2.53	46.02±2.10	48.69±2.16	70.68±6.30	72.40±6.52	71.04±5.97	71.18±6.14
LF-SVM	45.09±4.82	49.90±5.18	48.18±4.01	51.40±3.96	67.59±5.60	69.37±5.77	70.24±4.95	71.11±5.09
EmotionMeter	61.71±3.45	62.00±3.39	62.08±3.16	62.18±3.08	85.26±2.52	79.59±2.70	80.02±3.32	80.18±3.25
MMResLSTM	65.68±2.13	66.39±2.05	66.31±1.78	66.39±1.86	**86.78±2.56**	**87.03±2.55**	**86.55±1.78**	**87.13±2.25**
MMFN	70.42±3.14	71.56±3.66	70.37±4.02	72.42±3.97	**85.78±5.32**	85.65±5.61	84.27±6.37	**87.42±5.91**
HusFuse	67.45±3.23	68.64±3.16	67.85±2.67	68.08±2.53	80.48±1.58	80.77±1.71	81.26±1.38	81.42±1.48
HusLSTM	72.66±2.34	73.09±2.37	71.03±1.90	71.40±1.93	83.41±1.90	84.15±2.09	84.61±1.58	84.73±1.47
HusPair	**77.14±2.40**	**76.71±2.18**	**77.55±2.22**	**77.05±2.09**	**89.42±3.33**	**89.26±2.99**	**90.31±2.99**	**90.15±3.06**
Husformer	**79.64±1.52**	**79.87±1.54**	**79.94±2.18**	**80.44±2.25**	**90.67±2.20**	**90.74±2.29**	**91.33±1.59**	**91.35±1.67**

TABLE VI
PERFORMANCE OF DIFFERENT MODELS ON WESAD, RAW MOCAS, PREPROCESSED MOCAS, AND COGLOAD DATASETS IN TERMS OF AVERAGE MULTI-CLASS AVERAGE ACCURACY (Acc) AND MULTI-CLASS AVERAGE F1-SCORE ($F1$) WITH STAND DEVIATIONS. RESULTS OF OTHER MODELS THAT ARE WITHIN 5% OF THE *Husformer*'S PERFORMANCE ON Acc OR $F1$ ARE HIGHLIGHTED. ^{h}: HIGHER VALUES INDICATE BETTER PERFORMANCE.

Dataset	WESAD		Raw MOCAS		Preprocessed MOCAS		CogLoad	
Metric	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$
EF-SVM	42.46±4.34	44.39±4.08	51.48±4.39	51.63±5.00	62.73±4.91	61.87±4.19	41.67±3.80	47.52±3.14
LF-SVM	44.98±2.48	47.51±3.00	48.74±3.40	48.85±3.42	59.80±5.16	60.68±5.07	38.98±2.71	45.87±2.12
EmotionMeter	63.01±1.41	63.21±1.34	71.15±3.48	70.98±3.39	78.80±2.54	79.94±2.61	59.57±1.42	62.99±1.30
MMResLSTM	65.76±1.12	66.32±1.24	75.33±2.41	75.44±2.21	82.81±1.34	83.25±1.40	61.44±1.67	63.39±1.71
MMFN	64.68 ±11.12	59.21 ±12.05	65.38±8.26	70.71±9.83	81.17±6.17	83.54±5.16	65.50±1.17	65.93±1.42
HusFuse	68.77±1.56	68.48±1.31	70.65±2.36	71.22±2.39	78.00±2.10	78.81±1.86	58.49±0.68	57.65±0.83
HusLSTM	70.64±1.21	71.00±1.28	78.98±2.72	79.28±2.61	82.40±1.80	82.54±1.78	67.09±1.06	66.60±1.02
HusPair	73.57±1.72	73.77±2.13	82.12±1.83	82.46±1.63	**88.83±3.97**	**88.75±3.99**	65.11±3.07	66.55±3.25
Husformer	**78.68±2.05**	**79.51±2.28**	**87.37±2.40**	**87.47±2.55**	**90.09±2.25**	**90.17±2.17**	**74.06±2.48**	**74.93±2.77**

V. RESULTS AND ANALYSIS

A. Quantitative Measurements

1) Comparative Results with Baselines: Tables V and VI summarize the performance of our *Husformer* when compared with the five multi-modal baselines of human state recognition in terms of and with stand deviations during the experiments. From the comparative results, it is evident that our proposed *Husformer* consistently surpasses the other five state-of-the-art baselines in recognizing both human affective states (as seen on the DEAP and WESAD datasets) and cognitive load (as evidenced on the MOCAS and CogLoad datasets). Furthermore, this superior performance is maintained across all four datasets, each with its unique combination of modalities. Such consistency underscores the potential of the *Husformer* to serve as a robust backbone network for a broad spectrum of human state prediction tasks. Additionally, such performance enhancements are more evident on datasets without further feature engineering (raw DEAP, raw MOCAS, WESAD, and CogLoad datasets). This suggests that the *Husformer* can learn from raw multi-modal features series more efficiently, and thus is more applicable to real-world task scenarios where extensive feature crafting and alignment is quite impractical and expensive. These performance improvements mainly result from the following reasons:

Experimental results show that our *Husformer* significantly outperforms *EF-SVM* and *LF-SVM* regarding and across all datasets. This is reasonable since these two methods resort to a straightforward concatenation of multiple modalities either at the feature or decision level. Such an approach neglects the potential correlations among different modalities and may become susceptible to the curse of dimensionality, as pointed out by [51]. In contrast, *Husformer* seamlessly integrates unimodal features using a feed-forward process within its multi-layered cross-modal attention transformers, ensuring a thorough consideration of the synergistic interactions across various modalities. Furthermore, the disregard of cross-modal interactions by *EF-SVM* and *LF-SVM* also leads them to lag behind other baselines: *EmotionMeter* and *MMResLSTM*. This performance dip is evident with a roughly 10% decrease in both and . Such comparative outcomes underscore the pivotal role of recognizing and leveraging cross-modal correlations in multi-modal human state detection.

Additionally, the SVM architecture, foundational to both *EF-SVM* and *LF-SVM*, is known to falter in efficiency when handling vast datasets [52]. Its vulnerability to missing values, outliers, and noisy data means an over-dependence on meticulous data cleansing and judicious feature selection [53]. While certain research underscores the capability of *EF-SVM* and *LF-SVM* to estimate human states accurately with rigorous data preprocessing and feature engineering [21], it is logical to conclude that these methods may falter when managing relatively unprocessed multi-modal features.

The experimental results show that our proposed *Husformer* significantly outperforms *EmotionMeter* in terms of and

on all datasets. We believe this superior performance stems from the inherent limitations of *EmotionMeter*. While *EmotionMeter* does account for cross-modal interactions by implementing shared hidden layers for joint representation learning, its RBMs fall short in capturing crucial time-dependent cross-modal interactions. Given that temporal dynamics play a pivotal role in signals mirroring human states [41], this oversight likely undermines its efficacy.

On the other hand, *MMResLSTM* and *MMFN* outperform *EmotionMeter*, and are in close competition with our *Husformer* on the preprocessed DEAP dataset. This is attributed to their more advanced strategies in exploring cross-modal relations: *MMResLSTM* constructs LSTM layers that share the same weights for different modalities to learn shared representations, which can model the temporal cross-modal correlations effectively; and the *MMFN* computes the complementarity and importance within multiple modalities with dot-product attention.

However, on other datasets, particularly those with minimal feature engineering, *Husformer* consistently outshines both *MMResLSTM* and *MMFN* in terms of and . We believe this superiority arises from the greater effectiveness of the cross-modal attention mechanism within the *Husformer* in terms of capturing cross-modal dependencies. Unlike weight-sharing strategies in *MMResLSTM*, our approach fosters cross-modal interactions by prompting one modality to directly engage with the unimodal features of other modalities. This process identifies and leverages strongly complementary representational information to enhance the primary modality.

Moreover, when compared to the dot-product or weight attention techniques employed by *MMFN*, *Husformer*'s cross-modal attention mechanism facilitates more intricate and adaptable inter-modality interactions. It allows complementary information from various modalities to fluidly integrate into a single unimodal representation by fusion of Keys and Values under the condition of the unimodal Queries as described in (5). Additionally, in contrast to the dot-product attention used in *MMFN*, the cross-modal attention can process all positions (or tokens) within the feature sequences simultaneously, leading to a more comprehensive capture of long-range dependencies. As a consequence, *MMFN* may primarily learn short-term attention, potentially treating unrelated or noisy data as significant information. This distinction is markedly evident by the considerable deviations in and , especially observable on the WESAD dataset.

2) Comparative Results with Single Modality: Fig. 3 contrasts the performance of our *Husformer* with the highest recognition results achieved using a single modality for each dataset, benchmarked against the generally best-performing multi-modal fusion baseline, *MMResLSTM*, in terms of . Comprehensive classification results for every individual modality across datasets, considering both and , are detailed in Appendix B. As evidenced in Fig. 3, *Husformer* consistently surpasses the top single-modal-based recognition outcomes across all four datasets in terms of .

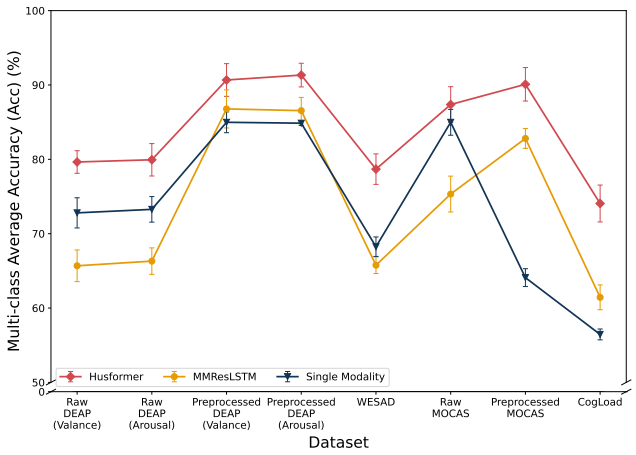


Fig. 3. Performance comparison of our proposed *Husformer* and the best-performing multi-modal fusion baseline, *MMResLSTM*, with the highest recognition result achieved using a single modality on each dataset in terms of multi-class average accuracy (Acc).

We believe this superior performance is because *Husformer* can effectively harness the benefits of multi-modal fusion in human state recognition. These advantages include extracting vital representation information by integrating metrics from multiple sources, which is not possible with a single source [54], and enhancing the signal quality by minimizing noise [6].

Interestingly, the *MMResLSTM*, despite its multi-modal fusion prowess that sets it ahead of other baselines, does not surpass the results of transformer networks focused solely on EEG-related modality when tested on both raw DEAP and preprocessed MOCAS datasets. This observation underscores the potency of the cross-modal attention and self-attention mechanisms in *Husformer* again. Specifically, these features in *Husformer* not only adeptly model intricate and extended complementary cross-modal interactions, but also accentuate the most relevant contextual representations, allowing *Husformer* to harness the benefits of multi-modal fusion more effectively.

3) Results of the Ablation Study: Tables V and VI present the performance of the *Husformer* in terms of and on each dataset, compared to three ablation models, the *HusFuse*, *HusPair*, and *HusLSTM*. From the results, we can observe the effectiveness of the cross-modal and self-attention module inside the *Husformer* as follows:

Effectiveness of the cross-modal attention module

Compared to the *HusFuse*, which removes the cross-modal attention module from the *Husformer*, the *Husformer* achieves an absolute improvement in terms of and . This improvement can be attributed to the fact that the *HusFuse* fuses low-level features from different modalities together using a simple concatenation operation, instead of utilizing a feed-forward process from cross-modal transformers. While the self-attention processes in the *HusFuse* can be viewed as a way to consider cross-model interactions, by relating entities of the sequence concatenated from low-level features of different modalities to contextual information to calculate the high-level fusion representation, it fails to provide direct complementary adaptions for features of one modality with those of other modalities during the fusion process. This result highlights the effectiveness of

TABLE VII
THE NUMBER OF PARAMETERS AND GPU MEMORY USAGE DURING TRAINING OF THE *Husformer* AND *Huspair* ON EACH DATASET. NOTE THAT BOTH MODELS HAD THE SAME BATCH SIZE DURING TRAINING. PARA: THE NUMBER OF PARAMETERS; MEM: GPU MEMORY USAGE.

Dataset	Raw DEAP		Preprocessed DEAP		WESAD		Raw MOCAS		Preprocessed MOCAS		CogLoad	
Metric	Para	Mem	Para	Mem	Para	Mem	Para	Mem	Para	Mem	Para	Mem
HusPair	2.90M	3253.24MB	2.92M	2908.66MB	3.90M	5416.90MB	6.21M	519.67MB	6.22M	531.80MB	3.12M	202.47MB
Husformer	0.63M	1999.56MB	0.66M	1773.59MB	0.71M	3084.35MB	0.74M	210.97MB	0.75M	220.53MB	0.72M	94.28MB

the proposed cross-modal attention-based fusion strategy compared to simple concatenation with self-attention.

Compared with the *HusPair*, which replaces the cross-modal attention module in the *Husformer* with the directional pairwise cross-modal attention in [27], the *Husformer* achieves an improvement in and . We argue that this is because the pairwise cross-modal attention in the *HusPair* can only consider the complementary interactions between a pair of modalities at once and thus ignores the coordination among more than two modalities. In contrast, our proposed cross-modal attention computes the complementary interactions between the low-level unimodal features of one target modality and the low-level fusion representation embedded with unimodal features of the other source modalities. This allows the model to consider the coordination across all modalities at the same time, hence considering more long-term and comprehensive cross-modal interactions. Moreover, it has been shown that the pairwise fusion approach can produce redundant fusion information that may serve as additional noise rather than effectual multi-modal features [55].

We can also observe that *HusPair* is quite comparable to the *Husformer* on the DEAP datasets and preprocessed MOCAS dataset. However, as presented in Table VII, the parameter number of the *HusPair* is about 4 8 times that of the *Husformer*. This is because the number of the pairwise cross-modal attention transformers in the *HusPair* increases exponentially with the increase in the number of modalities. Specifically, when applied to multi-modal fusion of modalities, the pairwise cross-modal attention requires cross-modal transformers, while ours only requires of them. For instance, on the WESAD dataset that contains six modalities, the *HusPair* requires cross-modal transformers, while our *Husformer* only requires of them. Such a high volume of parameters can result in slow convergence and high training difficulty. For example, on the preprocessed DEAP and the preprocessed MOCAS, where the *HusPair* gets its most competitive results, we empirically observe that the *Husformer* can converge faster to a lower loss of mean absolute error compared to the *HusPair* during the training process (see Fig. 4).

Moreover, we posit that the pairwise cross-modal attention mechanism in *HusPair*, which may be prone to over-parameterization, could lead to attention redundancy. This is particularly likely when the complexity of the features does not match that of the attention mechanism, resulting in the model mistaking noise or artifacts as meaningful attention components. For instance, in the CogLoad dataset, which features relatively lower feature-length (channel number) and dimension (sampling frequency) compared to other

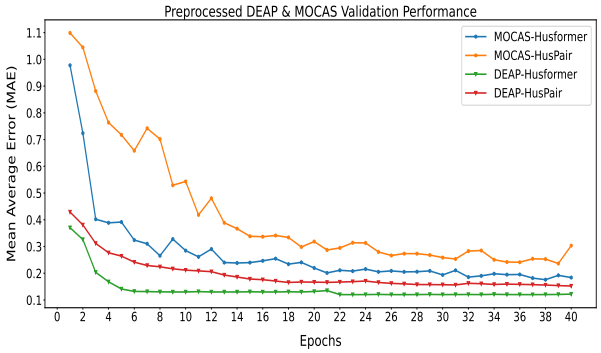


Fig. 4. Learning curves of *Husformer* when compared to *HusPair* on the preprocessed DEAP and MOCAS datasets in terms of validation set convergence.

datasets, *HusPair*'s performance notably diminishes. In fact, it lags behind *Husformer* by a margin of 10%, as evidenced in Tables V and VI. The high storage and computational demands of *HusPair* (refer to Table VII), coupled with its less effective performance on datasets with simpler modalities, restrict its practicality in diverse multi-modal human state recognition scenarios. Conversely, our *Husformer* not only achieves comparable but in some cases significantly superior performance, and it does so with substantially fewer parameters than *HusPair*. These findings underscore the efficiency and efficacy of our proposed cross-modal attention mechanism in contrast to the directional pairwise attention approach used in *HusPair*.

Effectiveness of the self-attention module

Compared with the *HusLSTM* that replaces the self-attention module in the *Husformer* with an LSTM layer, the *Husformer* achieves an absolute improvement on and . We believe that such an improvement results from the fact that the self-attention mechanism applied in the transformer network [30] enables the model to capture long-term temporal dependencies by considering the sequence consisting of all reinforced unimodal features as a whole. In contrast, the LSTM processes the sequence element by element, which can suffer from long-dependency issues [56]. Moreover, the self-attention mechanism can adaptively highlight meaningful contextual information while reducing useless ones by computing adaptive attention scores at a different position in the sequence, leading to a more effectual global representation of the human state. This result confirms the effectiveness of the self-attention module in the *Husformer*.

B. Qualitative Analysis

To demonstrate how the cross-modal attention and self-attention in the *Husformer* work when learning from multi-modal signals of the human state, we visualize the attention ac-

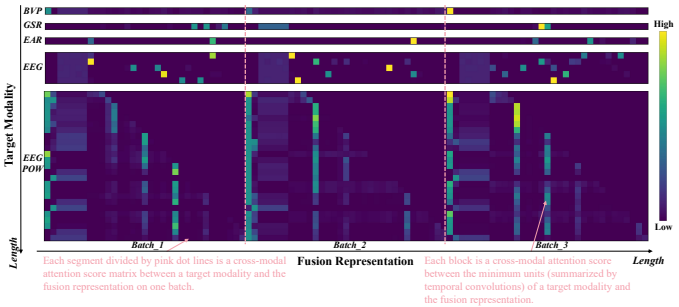


Fig. 5. Visualization of an example cross-modal attention weight group consisting of learned cross-modal attention matrices at the final layer of each cross-modal transformer within 3 batches during the training on the raw MOCAS dataset. Note that the cross-modal attention score matrix between a target low-level unimodal feature and the low-level fusion representation on one batch has the dimension of $L_M \times L_F$, i.e., the length of the target unimodal feature plus that of the fusion representation (BVP, GSR and EAR: 1×33; EEG: 5×33; EEG_POW: 25×33).

tivation for qualitative analysis. Fig. 5 shows an example cross-modal attention weight group consisting of learned cross-modal attention matrices at the final layer of each cross-modal transformer within 3 batches during the training on the raw MOCAS dataset. Note that the original cross-modal attention score matrix between a target low-level unimodal feature and the low-level fusion representation on one batch has the dimension of (see Fig. 2a). We can observe that the cross-modal attention has learned how to attend to positions revealing relevant and meaningful information across the target modality and source modalities embedded in the fusion representation without the requirement of feature alignment. For instance, higher cross-modal attention scores are assigned to some intersections of the BVP and GSR unimodal features and part of the EEG_POW unimodal features embedded in the later part (8-32) of the fusion representation. This shows that our cross-modal attention can reveal cross-modal contingencies that are inaccessible with manual feature alignment.

Furthermore, we can observe that the learned cross-modal attention is adaptive; i.e., different cross-modal attention patterns are learned between different target modalities and the fusion representation. Moreover, these patterns may differ for the same target modality across different patches. For instance, while the EEG modality is always encouraged to attend to the later section (8-32) of the fusion representation, i.e., the EEG_POW modality, the intersections assigned with higher cross-modal attention scores vary across different batches. However, despite the above adaptive differences, we can notice that some stable and consistent cross-modal attention patterns exist in different batches of the same target modality. For example, higher cross-modal attention scores are always assigned to the intersections of the EEG_POW unimodal features and the BVP, GSR and EEG unimodal features embedded in the front positions (1-7) in the fusion representation. These observations over the visualized cross-modal attention score matrices demonstrate that our proposed cross-modal attention module can capture and model an adaptive but relatively consistent and long-term pattern of cross-modal interactions.

Fig. 6a shows an example of the output of the self-attention transformer in the *Husformer*, namely the high-level fusion representation , within 3 batches during training

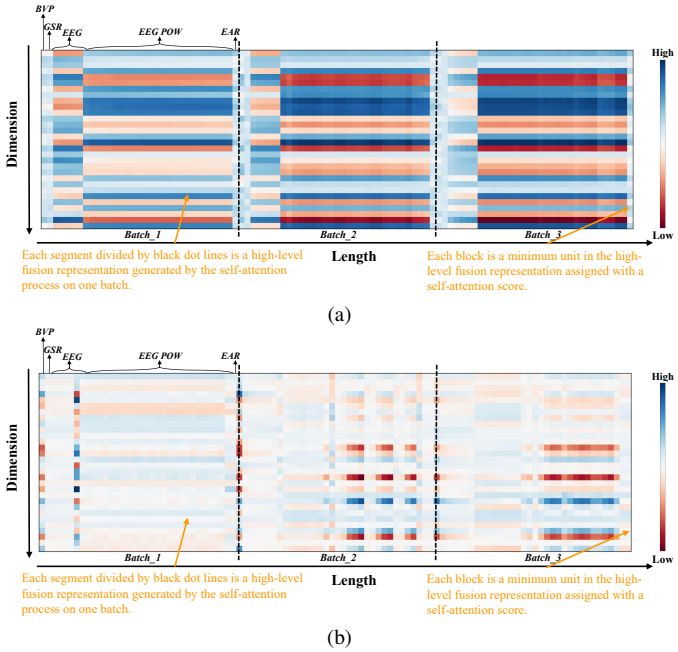


Fig. 6. Visualization of an example high-level fusion representation group generated by the final layer of the self-attention transformer in the (a) *Husformer* and (b) *HusFuse* within 3 batches during the training on the raw MOCAS dataset. Note that the high-level fusion representation produced on one batch has the dimension of $L_F \times D$, i.e., 33×30.

on the raw MOCAS dataset. We can observe that the self-attention has learned how to prioritize important contextual information for human state recognition and reduce insignificant ones by assigning high (blue) and low (red) self-attention scores to different positions in the fusion representation. Also, similar to the cross-modal attention, the learned self-attention is adaptive; i.e., different minimum units in the fusion representation are assigned with self-attention scores in different patterns, especially from the 'Dimension' axis. Meanwhile, we can notice that the learned self-attention patterns of the features of the same modality share many similarities across different batches, especially for EEG and EEG_POW modalities. These observations demonstrate that the self-attention in the *Husformer* can highlight effectual contextual features of the human state and diminish ineffectual ones in the fusion representation with an adaptive while relatively steady pattern.

Meanwhile, Fig. 6b depicts an example of the high-level fusion representation output of the self-attention transformer in the *HusFuse*. Note that the inputs of the self-attention transformers in the *Husformer* and *HusFuse* are concatenated by unimodal features of each modality within the same fragments. The only difference is that the unimodal features in the input of the *Husformer* are reinforced by cross-modal attention transformers while those in the *HusFuse* are not. Comparing the different self-attention patterns assigned to the same fragments of unimodal features in the fusion representation as illustrated in Figs. 6a and 6b, we can notice that without the cross-modal attention modeling the cross-modal interactions, the self-attention learned in the *HusFuse* is less efficient. That is, only few of high (blue) and low (red) self-attention scores are assigned to features in the fusion representation, leading to insufficient prominence of critical contextual information and diminishing of unimportant information respectively. We

can also notice that the self-attention is less consistent. That is, no stable self-attention patterns are shown on features of the same modality across batches, especially between *Batch_1* and *Batch_3*. Such differences demonstrate that the reinforcements for unimodal features from the cross-modal attention in the *Husformer* can help the self-attention highlight critical contextual information in the fusion representation with a more efficient and consistent pattern.

C. Real-world Experiment

Furthermore, we applied the *Husformer* to a real-world CCTV monitoring task, similar to the procedure used in the creation of the MOCAS dataset. In this task, 32 participants were asked to identify abnormal objects from a video streaming from a multi-robot system. The monitoring tasks were differentiated into three levels based on the number of cameras and robot speed, with the goal of simulating varying cognitive workloads. Over the course of the study, each participant engaged in eight separate tasks. Participants were equipped with two wearable sensors: the E4 wristband and the Emotiv Headset. These devices collected physiological modalities such as BVP, GSR, EEG, and EEG_POW. Additionally, a Realsense D350 camera was used to capture the behavioral modality of EAR. All signals were gathered in real time with a sampling rate of 100 Hz. After each monitoring task, participants reported their subjective cognitive workload as the annotations via the NASA_TLX survey.

Trained using the MOCAS dataset, the *Husformer* was able to predict the cognitive workload (categorized as low, medium, or high) of multiple participants at 100 Hz on an NVIDIA 3060 GPU. The mean predicted outputs throughout each monitoring task were considered the final objective cognitive workload output. Across the total of 256 task periods involving the 32 new human subjects, the *Husformer* achieved an average accuracy rate of 70.31%. Furthermore, by leveraging the objectively predicted cognitive loads from the *Husformer*, the overall efficiency of the human-robot team was enhanced by dynamically adjusting the workload assigned to each human operator. More details can be found in [57].

VI. Conclusion and Future Work

In this paper, we proposed the *Husformer*, an end-to-end multi-modal transformer framework for recognizing human affective state or cognitive load from multiple modalities. The *Husformer* fuses modalities with adaptive and sufficient cross-modal interactions, enabling one modality to attend to features of other modalities where strong cross-modal relevance exists. It also adaptively highlights important contextual information in the fusion representation. These two attention mechanisms, operating at the inter-modal and fusion representation levels, enable our model to efficiently learn from multi-modal features, eliminating the need for extensive feature engineering and alignment required in previous works. Our experimental results on four public benchmark multi-modal datasets of human emotion and cognitive load demonstrated the effectiveness of the proposed *Husformer* for general human state recognition, outperforming five other state-of-the-art multi-modal-based baselines and demonstrating enhanced performance over using a single modality. Additionally, our ablation study highlighted the effectiveness of two key components in the *Husformer*: the cross-modal attention and self-attention modules.

Despite the promising results of the *Husformer*, we openly acknowledge certain limitations, which, in turn, pave the way for stimulating exciting future work. One limitation of the *Husformer* is its lack of adaptability to individual user variations, potentially reducing its efficacy with new human subjects. This limitation is manifested as a perceptible performance decline from the MOCAS dataset to real-world experiments. For future enhancements, it is imperative to integrate contextual information, such as user-specific data encompassing personality and demographic aspects. This inclusion will potentially steer the attention mechanism more adeptly, bolstering the model's versatility and efficiency across a diverse array of users.

Another area of concern is the potential inefficiency of the multimodal transformer network in situations of missing modalities during testing or real-world deployment, where some modalities might be either unavailable or substantially disrupted. This challenge is not exclusive to our model, but is also observed in other state-of-the-art multi-modal transformer networks, as depicted in [58]. Subsequent studies should place emphasis on enhancing the model's robustness against the challenges posed by missing or incomplete modal data. Current advancements in the domain of missing modality learning, as depicted in [59], [60], will be instrumental.

Moreover, the *Husformer* may tend toward overfitting when grappling with highly unbalanced data. To counteract this, we have meticulously designed and integrated a focal loss function within our model. Nevertheless, continued research is crucial to further augment the *Husformer* robustness against imbalanced datasets. Prospective research should delve into exploring and incorporating more advanced techniques or algorithms adept at managing imbalance. This exploration may encompass the integration of sophisticated sampling methods, cost-sensitive learning, or other pertinent strategies, enhancing the model performance and stability in confronting unbalanced data. Additionally, while the *Husformer* demonstrates adeptness in predicting cognitive load and emotion independently, it currently lacks the capacity for simultaneous predictions. Future advancements anticipate the availability of comprehensive datasets, annotated with both cognitive load and emotion metrics. This development will necessitate the incorporation of multi-task learning frameworks into the *Husformer*, enhancing its functionality to effectively predict and analyze multiple human state dimensions concurrently.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1846221. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] S. Aldini, A. K. Singh, D. Leong, Y.-K. Wang, M. G. Carmichael, D. Liu, and C.-T. Lin, "Detection and estimation of cognitive conflict during physical human-robot collaboration," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[2] M. Lagomarsino, M. Lorenzini, P. Balatti, E. De Momi, and A. Ajoudani, "Pick the right co-worker: Online assessment of cognitive ergonomics in human-robot collaborative assembly," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[3] D. D. Chakladar, S. Datta, P. P. Roy, and A. Vinod, "Cognitive workload estimation using variational auto encoder & attention-based deep model," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[4] Y. Yan, X. Wu, C. Li, Y. He, Z. Zhang, H. Li, A. Li, and L. Wang, "Topological eeg nonlinear dynamics analysis for emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[5] R. Wang, D. Zhao, and B.-C. Min, "Initial task allocation for multi-human multi-robot teams with attention-based deep reinforcement learning," *arXiv preprint arXiv:2303.02486*, 2023.

[6] E. Debie, R. F. Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, and H. A. Abbass, "Multimodal fusion for objective assessment of cognitive workload: a review," *IEEE transactions on cybernetics*, vol. 51, no. 3, pp. 1542–1555, 2019.

[7] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, and D. Zhang, "Cognitive workload recognition using eeg signals and machine learning: a review," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[8] Z.-T. Liu, S.-J. Hu, J. She, Z. Yang, and X. Xu, "Electroencephalogram emotion recognition using combined features in variational mode decomposition domain," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[9] Z. He, Z. Li, F. Yang, L. Wang, J. Li, C. Zhou, and J. Pan, "Advances in multimodal emotion recognition based on brain–computer interfaces," *Brain sciences*, vol. 10, no. 10, p. 687, 2020.

[10] Y. Peng, F. Qin, W. Kong, Y. Ge, F. Nie, and A. Cichocki, "Gfil: A unified framework for the importance analysis of features, frequency bands, and channels in eeg-based emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 935–947, 2021.

[11] J. W. Li, S. Barma, S. H. Pun, M. I. Vai, and P. U. Mak, "Emotion recognition based on eeg brain rhythm sequencing technique," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 1, pp. 163–174, 2022.

[12] J.-L. Kruger, S. Doherty, W. Fox, and P. De Lissa, "Multimodal measurement of cognitive load during subtitle processing," *Innovation and expansion in translation process research*, vol. 267, 2018.

[13] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and eeg for multimodal emotion recognition," *Computational intelligence and neuroscience*, vol. 2017, 2017.

[14] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep neural networks," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 811–819.

[15] J. Singh and R. Gill, "Multimodal emotion recognition system using machine learning and psychological signals: A review," *Soft Computing: Theories and Applications: Proceedings of SoCTA 2020, Volume 1*, pp. 657–666, 2022.

[16] T. Horii, Y. Nagai, and M. Asada, "Modeling development of multimodal emotion perception guided by tactile dominance and perceptual improvement," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 762–775, 2018.

[17] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[18] J.-L. Qiu, W. Liu, and B.-L. Lu, "Multi-view emotion recognition using deep canonical correlation analysis," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 221–231.

[19] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[20] S. Liu, P. Gao, Y. Li, W. Fu, and W. Ding, "Multi-modal fusion network with complementarity and importance for emotion recognition," *Information Sciences*, vol. 619, pp. 679–694, 2023.

[21] T. Zhou, J. S. Cha, G. Gonzalez, J. P. Wachs, C. P. Sundaram, and D. Yu, "Multimodal physiological signals for workload prediction in robot-assisted surgery," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 2, pp. 1–26, 2020.

[22] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[23] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[24] W. Jo, R. Wang, S. Sun, R. K. Senthilkumaran, D. Foti, and B.-C. Min, "MOCAS: A Multimodal Dataset for Objective Cognitive Workload Assessment on Simultaneous Tasks," Aug. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7023242

[25] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, and V. Pejović, "Datasets for cognitive load inference using wearable sensors and psychological traits," *Applied Sciences*, vol. 10, no. 11, p. 3843, 2020.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[27] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[28] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[29] J. Zhou, K. Yu, F. Chen, Y. Wang, and S. Z. Arshad, "Multimodal behavioral and physiological signals as indicators of cognitive load," in *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, 2018, pp. 287–329.

[30] 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.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[32] W. Jo, R. Wang, S. Sun, R. K. Senthilkumaran, D. Foti, and B.-C. Min, "Mocas: A multimodal dataset for objective cognitive workload assessment on simultaneous tasks," *arXiv preprint arXiv:2210.03065*, 2022.

[33] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[34] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior research methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[35] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, 2010, pp. 301–310.

[36] M. A. Hogervorst, A.-M. Brouwer, and J. B. Van Erp, "Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload," *Frontiers in neuroscience*, vol. 8, p. 322, 2014.

[37] P. Zhang, X. Wang, J. Chen, and W. You, "Feature weight driven interactive mutual information modeling for heterogeneous bio-signal fusion to estimate mental workload," *Sensors*, vol. 17, no. 10, p. 2315, 2017.

[38] F. Putze, J.-P. Jarvis, and T. Schultz, "Multimodal recognition of cognitive workload for multitasking in the car," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3748–3751.

[39] T. Zhou and J. P. Wachs, "Early prediction for physical human robot collaboration in the operating room," *Autonomous Robots*, vol. 42, no. 5, pp. 977–995, 2018.

[40] G. Shafer, "Dempster-shafer theory," *Encyclopedia of artificial intelligence*, vol. 1, pp. 330–331, 1992.

[41] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual lstm network," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 176–183.

[42] A. Graves, "Long short-term memory," in *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, pp. 37–45.

[43] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[44] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2018.

[45] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, 2020.

[46] J. Sun, J. Xie, and H. Zhou, "Eeg classification with transformer-based models," in *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, 2021, pp. 92–93.

[47] S. J. Russell, *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[48] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: with applications in R*. Spinger, 2013.

[49] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[50] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1439–1449.

[51] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA transactions on signal and information processing*, vol. 3, 2014.

[52] K. Hou, G. Shao, H. Wang, L. Zheng, Q. Zhang, S. Wu, and W. Hu, "Research on practical power system stability analysis algorithm based on modified svm," *Protection and Control of Modern Power Systems*, vol. 3, no. 1, pp. 1–7, 2018.

[53] Y. Wu and Y. Liu, "Robust truncated hinge loss support vector machines," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 974–983, 2007.

[54] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[55] Y. Bian, J. Huang, X. Cai, J. Yuan, and K. Church, "On attention redundancy: A comprehensive study," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 930–945.

[56] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and lstm encoder decoder models for asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 8–15.

[57] W. Jo, R. Wang, B. Yang, D. Foti, M. Rastgaar, and B.-C. Min, "Affective workload allocation for multi-human multi-robot teams," *arXiv preprint arXiv:2303.10465*, 2023.

[58] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.

[59] S. Karimijafarbigloo, R. Azad, A. Kazerouni, S. Ebadollahi, and D. Merhof, "Mmcformer: Missing modality compensation transformer for brain tumor segmentation," in *Medical Imaging with Deep Learning*, 2023.

[60] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 878–15 887.

APPENDIX A
HYPER-PARAMETERS OF THE *Husformer*

To optimize the hyper-parameters of the *Husformer*, we conducted a grid search. This process involved altering one hyper-parameter at a time, keeping others fixed, and observing the corresponding model performance on a validation set, distinct from the training and test sets. Specifically, we employed the SGD optimization scheme, experimenting with learning rates from 1e-5 to 1e-2, expanded to 2e-2 on the DAEP dataset. During the model training process, we varied the epochs from 40 to 160, halting training if the validation loss failed to decrease for 15 consecutive epochs. Considering the data shape and computation speed, we tested a batch size range from 16 to 2048 for each dataset. For transformer parameters, we examined the number of transformer layers between 2 to 10 and found that 5 layers offered an optimal balance between model complexity and predictive capability. We also tested the number of attention heads from 3 to 10 based on our dataset and the complexity of the interactions we aimed to capture. For the parameters of focal loss, we conducted a similar exploration, testing alpha values from 0.1 to 0.9 and gamma from 0.5 to 5. Furthermore, we experimented with the output dimension of the temporal convolution layer from 10 to 60, ultimately settling on a feature dimension of d=30 for feature fusion. The selected hyperparameters for the *Husformer* are outlined in Table VIII.

TABLE VIII
HYPER-PARAMETERS OF THE *Husformer* UTILIZED FOR EACH EXPERIMENT

Parameter name	Raw DEAP	Preprocessed DEAP	WESAD	Raw MOCAS	Preprocessed MOCAS	Cogload
Batch Size	1024	1024	512	64	128	1024
Initial Learning Rate	2e-3	2e-3	1e-3	1e-3	1e-3	1e-3
Optimizer	SGD	SGD	SGD	SGD	SGD	SGD
Transformer Hidden Unit Size	40	40	40	40	40	40
Crossmodal Attention Heads	3	3	3	5	5	3
Crossmodal Attention Block Dropout	0.1	0.1	0.05	0.05	0.05	0.1
Output Dropout	0.1	0.1	0.1	0.1	0.1	0.1
Focal Loss α_c	[0.1,0.1,0.8]	[0.15,0.05,0.8]	[0.4,0.3,0.3]	[0.2,0.1,0.7]	[0.15,0.15,0.7]	[0.1,0.1,0.8]
Focal Loss γ	3	3	2	3	3	2
Epochs	120	120	80	100	100	120

APPENDIX B
CLASSIFICATION RESULTS USING EACH SINGLE MODALITY

TABLE IX
BEST PERFORMING CLASSIFICATION RESULTS OF USING SINGLE MODALITY ON THE RAW DEAP AND PREPROCESSED DEAP DATASET IN TERMS OF MULTI-CLASS AVERAGE ACCURACY (Acc) AND MULTI-CLASS AVERAGE F1-SCORE ($F1$) WITH STAND DEVIATIONS. ALL BEST-PERFORMING RESULTS ARE OBTAINED WITH THE TRANSFORMER NETWORK.

Dataset	Raw DEAP				Preprocessed DEAP			
Criteria	Valence		Arousal		Valence		Arousal	
Metric	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$
EEG	72.80±2.03	72.93±2.20	73.27±1.72	73.71±1.93	84.98±1.40	85.09±1.42	84.86±0.30	84.86±0.80
EMG	64.66±2.38	64.71±2.33	65.32±2.24	65.66±1.83	76.69±1.04	76.71±4.02	74.09±0.89	74.07±0.86
EOG	46.99±1.93	45.98±1.83	48.27±1.57	50.97±1.44	62.76±1.85	65.57±1.45	64.42±0.44	65.43±0.39
GSR	45.18±2.56	45.63±2.38	46.70±2.33	48.07±2.17	61.65±1.16	59.74±0.97	62.75±0.91	57.28±0.88

TABLE X
BEST PERFORMING CLASSIFICATION RESULTS OF USING SINGLE MODALITY ON THE ALL DATASETS IN TERMS OF MULTI-CLASS AVERAGE ACCURACY (Acc) AND MULTI-CLASS AVERAGE F1-SCORE ($F1$) WITH STANDARD DEVIATIONS; ♣: CLASSIFICATION WITH TRANSFORMER, AND ♠: CLASSIFICATION WITH GCN.

Dataset	Raw MOCAS		Preprocessed MOCAS		WESAD		CogLoad	
Metric	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$	$Acc(\%)^h$	$F1(\%)^h$
EEG	34.17±0.57♣	34.01±0.53♣	44.80±0.37♣	46.21±0.37♣	-	-	-	-
EEG_POW	68.25±1.31♣	67.87±1.45♣	84.98±1.74♣	84.90±1.72♣	-	-	-	-
GSR	33.70±0.61♣	35.59±0.64♣	43.44±0.57♣	46.98±0.53♣	-	-	56.45±0.73♣	57.52±0.80♣
BVP	42.55±1.63♣	43.49±1.71♣	71.18±2.03♣	71.14±2.06♣	60.75±0.95♣	61.27±0.99♣	-	-
EAR	47.34±3.12♣	49.29±2.29♣	51.46±0.42♣	48.37±0.45♣	-	-	-	-
EMG	-	-	-	-	52.71±0.46♣	55.62±0.54♣	-	-
EDA	-	-	-	-	53.85±0.77♣	56.45±0.53♣	-	-
RESP	-	-	-	-	64.09±1.20♣	65.85±1.04♣	-	-
HR	-	-	-	-	-	-	30.54±1.13♣	30.03±1.21♣
RR	-	-	-	-	-	-	39.88±2.32♣	42.58±2.47♣
ACC	-	-	-	-	-	-	34.59±1.81♣	33.35±1.53♣