A Dense Convolutional Bi-Mamba Framework for EEG-Based Emotion Recognition

Mingya Zhang[1,] **Yuqian Zhuang**[1,] **Liang Wang**[1,*] **Yiyuan Ge**[2] **Zhihao Chen**[2] **Xianping Tao**[1]
[1] Nanjing University [2] Beijing Information Science and Technology University

Abstract

In recent times, emotion recognition based on electroencephalograms (EEGs) has found extensive applications. Although numerous approaches leveraging CNN and Transformer have been put forward for automatic emotion recognition and have achieved commendable performance, several challenges remain: (1) Transformer-based models are proficient at capturing long-term dependencies within EEG signals. However, their quadratic computational complexity poses a significant hurdle. (2) Models that combine Transformers with convolutional neural networks (CNNs) often fail to effectively capture the coarse-to-fine temporal dynamics of EEG signals. State Space Models (SSMs), exemplified by Mamba, have emerged as a promising solution. They not only showcase outstanding capabilities in modeling long-range interactions but also maintain a linear computational complexity, which is highly advantageous. To address these challenges head-on, we introduce Emotion-Mamba, an innovative framework designed specifically for EEG-based emotion recognition. The proposed framework initiates the process by employing the CNN Encoder to extract information from both the temporal and spatial dimensions of EEG signals. Subsequently, the extracted feature information is relayed to the Hierarchical Coarse-to-Fine Bi-Mamba (HBM) block, which is adept at efficiently processing these features. Furthermore, a Dense Temporal Fusion (DTF) module has been incorporated. This module capitalizes on the multi-level, purified temporal information sourced from CNN Encoder and HBM blocks, with the aim of bolstering decoding accuracy. We conduct comprehensive evaluations of Emotion-Mamba using the SEED and SEED-V datasets. The experimental findings unequivocally demonstrate that our proposed approach surpasses the existing state-of-the-art methods.

Keywords: EEG, Emotion Recognition, Mamba

Introduction

Brain-computer interface (BCI) technology facilitates direct communication between the brain and machines using electroencephalography (EEG) (Lotte & Guan, 2010). A standard BCI system usually consists of four key components: data acquisition, pre-processing, classification, and feedback (Lotte & Guan, 2010). BCIs are employed in various practical applications, such as stroke rehabilitation (Foong et al., 2019), sleep stage detection (Jia, Liang, et al., 2024; Jia, Wang, et al., 2024), and emotion regulation in mental health treatments (Zotev et al., 2020). Emotion represents an individual's physiological arousal and the cognitive response that adapts to this state (Schachter & Singer, 1962). In recent years, the study of automatic emotion recognition has garnered significant attention. Currently, two main types of signals are utilized in emotion recognition tasks: overt behavioral signals and human physiological signals (Dzedzickis et al., 2020). Emotion, as a complex cognitive process, has been shown to result from the coordinated activity of the cerebral cortex and subcortical nerves (Malfliet et al., 2017). Consequently, physiological signals represented by electroencephalograms (EEGs) possess an inherent advantage in emotion recognition tasks. An EEG signal is a type of human physiological electrical signal with high temporal resolution (Burle et al., 2015). Previous studies have also focused on the extraction of temporal features from EEGs, but these features tend to be considered from a single scale, local or global. For example, convolutional neural networks (CNNs) (Ozdemir et al., 2021) or long short-term memory (LSTM) networks (Feng et al., 2022) have been used. In emotional EEG signals, there is a strong relationship between adjacent temporal points, and a high degree of continuity (Mitchell, 2021). Additionally, there is a correlation between distant temporal points and similar neural patterns and representations (Riberto et al., 2022).

A variety of deep learning techniques have emerged for interpreting brain activities from EEG signals (Ding, Robinson, Zhang, et al., 2023; Lawhern et al., 2018; Schirrmeister et al., 2017). These techniques generally fall into two categories: those that rely on manually extracted features and those that use EEG data directly as input (Ding, Robinson, Zhang, et al., 2023). The first category involves the extraction of diverse features from EEG signals to feed into neural networks (Ding & Guan, 2023; T. Song et al., 2020). On the other hand, CNN-based approaches utilize automatic feature extraction and often treat EEG as a 2-D time series (Ding, Robinson, Zhang, et al., 2023; Lawhern et al., 2018; Schirrmeister et al., 2017). Since EEG data naturally forms a graph structure, with electrodes as nodes and connections based on spatial proximity, functional connectivity, or learned associations, graph-based methods have become increasingly popular (Ding & Guan, 2023; Ding, Robinson, Tong, et al., 2023). In addition to CNNs, transformer-based neural architectures have garnered considerable interest in the BCI field due to their ability to capture global dependencies (Y. Song et al., 2023; Vaswani et al., 2017). Typically, previous studies have employed a CNN-Transformer architecture, where the CNN component acts as an adaptive feature encoder to preprocess EEG data, and the transformer component then identifies long-range temporal features (Lee & Lee, 2022; Xie et al., 2022).

Recent advancements in State Space Models (SSMs), par-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

ticularly Structured SSMs (S4), provide an effective solution due to their proficiency in handling long sequences. e.g., Mamba (Gu & Dao, 2023). The Mamba model augments S4 with a selective mechanism and hardware optimization, demonstrating outstanding performance in dense data domains. The success of Mamba in handling Computer Vision and NLP tasks has influenced more research into various feature processing tasks based on Mamba, especially in its ability to construct long-sequence models. We believe that BCI EEG processing is particularly suitable for modeling with Mamba. Although previous works can capture either fine-grained (short-period) or coarse-grained (global) temporal dependencies within each layer, they have not explicitly captured both coarse- and fine-grained temporal dynamics within the Transformer layers, which may limit the full utilization of EEG signals' long-short period temporal dynamics (Xie et al., 2022).

To mitigate the above-mentioned issues and enhance the perception of temporal dynamics in EEG data, we introduce Emotion-Mamba, a novel dense convolutional Mamba framework. We propose a novel Hierarchical Coarse-to-Fine Bi-Mamba (HBM) block, It is used to meet the requirements of long distance modeling and reduces the computational complexity. Built upon a CNN-based feature encoder comprising collaborative temporal and spatial convolutional layers, HBM concurrently captures coarse- and fine-grained temporal dynamics information. These representations are then adaptively fused with the coarse-grained temporal representations encoded by the Bi-Mamba, thereby providing more discriminative long- and short-term temporal information. Furthermore, to efficiently utilize multi-level temporal information from intermediate HBM layers, we have designed a Dense Temporal Fusion (DTF) module in Emotion-Mamba. This module enables the dense transmission of multi-level representations from HBM layers to the final representation.

In summary, the contributions of this work are summarized as follows:

- We propose Emotion-Mamba, a dense convolutional Bi-Mamba model based on State Space Models (SSMs), which is designed for EEG emotion analysis tasks.
- We propose a Hierarchical coarse-to-fine Bi-Mamba (HBM) module, which is designed to effectively encode the coarse-to-fine temporal dynamics within EEG data.
- We designe a Dense Temporal information Fusion (DTF) module, which is used to emphasize the importance of EEG information in the temporal dimension during the encoding process and significantly improve the performance of the model.
- Numerous experiments have demonstrated that our Emotion-Mamba has achieved state-of-the-art (SOTA) performance on public emotion datasets.

Related Work

Emotion recognition based on traditional machine learning requires the human extraction of EEG features and the design of classifiers. The commonly used features include power spectral density (PSD), differential entropy (DE) and rational asymmetry (RASM) (X. Li et al., 2022). In terms of choosing classifiers, support vector machine (SVM) (Zhao et al., 2019) and XGBoost (Xefteris et al., 2022), among others, are widely used. However, machine learning methods for processing raw data are limited (LeCun et al., 2015).

CNNs for EEG

Shen et al. (Shen et al., 2020) combined CNNs and RNNs to extract frequency, temporal and spatial features of EEGs, and achieved 94.74% accuracy on the SEED. However, the size of the convolutional kernel—a large kernel limits the extraction of deep information, whereas a small kernel limits the perceptual field of view—tends to be the limiting factor for CNNs (He et al., 2019). Schirrmeister et al. (Schirrmeister et al., 2017) introduced DeepConvNet, which incorporates a two-stage spatial and temporal convolution layer to facilitate EEG feature extraction and classification. Similarly, Lawhern et al. (Lawhern et al., 2018) developed EEGNet, using depth-wise convolution with a kernel size of (n, 1) to capture spatial features. Building on these methods, TSception (Ding, Robinson, Zhang, et al., 2023) employs multi-scale convolutional kernels to decode temporal dynamics and asymmetric spatial activations within EEG signals. However, the relatively short length of these 1-D CNN kernels in the temporal dimension limits their ability to capture long-term temporal patterns effectively.

Transformers for EEG

The emergence of an attention mechanism effectively alleviates the problems mentioned above. Tao et al. (Tao et al., 2020) used a CNN incorporating channel-wise attention to extract more discriminative spatial information and explored temporal relationships via RNNs. A transformer based on a self-attention mechanism has the inherent ability to perceive global dependencies (Vaswani et al., 2017). Song et al. (Y. Song et al., 2022) proposed the EEG Conformer, a convolutional transformer model used to extract EEG temporal features. However, most existing convolutional Transformers, which typically leverage CNNs for shallow feature extraction followed by Transformer blocks, (Ding et al., 2024) proposed the EEG Deformer, which successfully learns both coarse- and fine-grained temporal patterns in EEG signals and fuses multi-level temporal information across different layers. However, the aforementioned work is all based on Transformers, and multi-stage Transformers further increase computational complexity and demand high performance. Moreover, we believe that in the initial feature extraction stage, directly performing spatial convolution on temporal information does not fully utilize the temporal features of EEG.

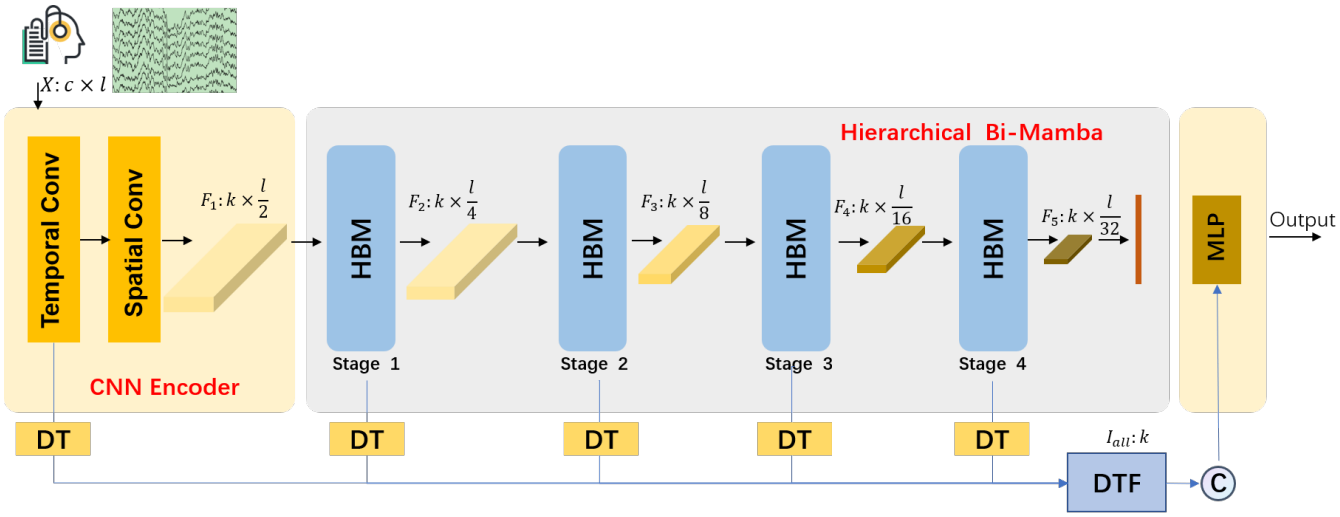

Figure 1: Emotion-Mamba consists of three main parts: (1) Convolution feature encoder, (2) Hierarchical Bi-Mamba (HBM), and (3) Dense Temporal information Fusion (DTF) with Dense Temporal (DT) units .

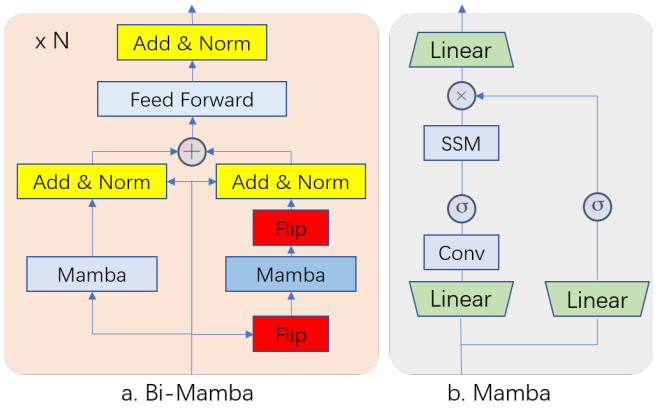

Figure 2: The architecture of (a) Bi-Mamba encoder and (b) Mamba block.

SSM and Mamba

State Space Models (SSMs) (Gu et al., 2021, 2022; Gupta et al., 2022) provide an efficient alternative to Transformers by capturing complex temporal dependencies with linear computational complexity, making them suitable for processing long-duration EEG sequences. Mamba further optimizes SSMs through a hardware-aware design and dynamic time-step selection strategy, improving computational efficiency and ensuring the model focuses on the most relevant EEG segments (Gu & Dao, 2023). However, both SSMs and Mamba still struggle with capturing bidirectional dependencies, which are critical for emotion recognition. The Bi-Mamba mechanism was specifically developed to overcome these limitations (Ruan & Xiang, 2024). By processing EEG data bidirectionally, BiMamba simultaneously captures past and future information, allowing for a more accurate representation , particularly in long sequences. This dual processing mechanism significantly improves the accuracy of emotion classification without increasing computational overhead, making it a superior solution for addressing the limitations of Transformers and traditional state space models. The integration of BiMamba into our proposed model allows it to effectively capture both short- and long-term dependencies, providing a scalable and efficient solution for emotion recognition.

Methods

The network architecture is shown in Figure 1. Emotion-Mamba consists of three main components: (1) Convolution feature encoder, (2) Hierarchical Bi-Mamba (HBM), and (3) Dense Temporal information Fusion (DTF). Emotion-Mamba utilizes the CNN feature encoder to adaptively encode temporal and spatial features, which are then set as input into the following HBM blocks to extract the temporal dynamics that happen in different timescales in EEG signals. To effectively perceive the critical multi-level temporal information, the features generated from CNN encoder and each HBM block are adaptively fused via Dense Temporal (DT) units and Dense Temporal information Fusion (DTF) module. Below, we present the details for each of them.

Preliminaries

In contemporary SSM-based models, namely, Structured State Space Sequence Models (S4) and Mamba (Gu & Dao, 2023; Liu et al., 2024; Ruan & Xiang, 2024), both depend on a traditional continuous system that maps a one-dimensional input function or sequence, represented as $x(t) \in R$, through intermediary implicit states $h(t) \in R^N$ to an output $y(t) \in R$. This process can be depicted as a linear Ordinary Differential Equation (ODE):

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) \end{aligned} \tag{1}$$

where $\mathbf{A} \in R^{N\times N}$ represents the state matrix, while $\mathbf{B} \in R^{N\times 1}$ and $\mathbf{C} \in R^{N\times 1}$ denote the projection parameters.

S4 and Mamba discretize this continuous system to adapt it better for deep learning contexts. Specifically, they incorporate a timescale parameter Δ and convert $\mathbf{A}$ and $\mathbf{B}$ into discrete parameters $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ using a consistent discretization rule. The zero-order hold (ZOH) is typically utilized as the discretization rule and can be outlined as follows:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}) \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A})-\mathbf{I})\cdot\Delta\mathbf{B} \end{aligned} \tag{2}$$

Following discretization, SSM-based models can be calculated in two distinct methods: linear recurrence or global convolution, which are denoted as equations (3) and (4), respectively.

$$\begin{aligned} h'(t) &= \overline{\mathbf{A}}h(t)+\overline{\mathbf{B}}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \tag{3}$$

$$\begin{aligned} \overline{K} &= \left(\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, ..., \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}\right) \\ y &= x * \overline{\mathbf{K}} \end{aligned} \tag{4}$$

where $\overline{\mathbf{K}} \in R^{L}$ represents a structured convolutional kernel, and L denotes the length of the input sequence x.

CNN Encoder

To capture the temporal and spatial information of EEG, we utilize a CNN-based feature encoder. CNNs along temporal and spatial dimensions are commonly used as feature extractors for EEG signals (Lawhern et al., 2018). In our proposed Emotion-Mamba, a two-layer CNN is adopted as the shallow feature encoder. The architecture of the CNN encoder begins with temporal and spatial CNN layers, followed by batch normalization to mitigate the covariate shift issue.

Analyze an EEG sample, denote it as $X \in R^{c\times l}$, where c represents the number of EEG channels and l denotes the number of data points along the time dimension. Drawing from neurophysiological insights that the brain exhibits microstates lasting approximately 100 ms, the temporal CNN kernels are set to a size of $(1, 0.1\times f_s))$, with f_s being the sampling rate of the EEG. To capture spatial information from EEG data, a spatial CNN kernel with dimensions of $(c,1)$ is employed, as recommended in reference (Lawhern et al., 2018), where c corresponds to the number of EEG channels. Weight normalization is applied in accordance with the method outlined in reference (Mane et al., 2020). The quantity of CNN kernels is represented by k. Subsequent to activation via the ELU function, the extracted features undergo max-pooling at every two data points without overlap. This step can be formularized as:

$$F = Re\left(MaxPool\left(ELU\left(BN\left(CNN\left(X\right)\right)\right)\right)\right) \tag{5}$$

where $Re(.)$ is the rearrange operation and $BN(.)$ is the batch normalization operation.

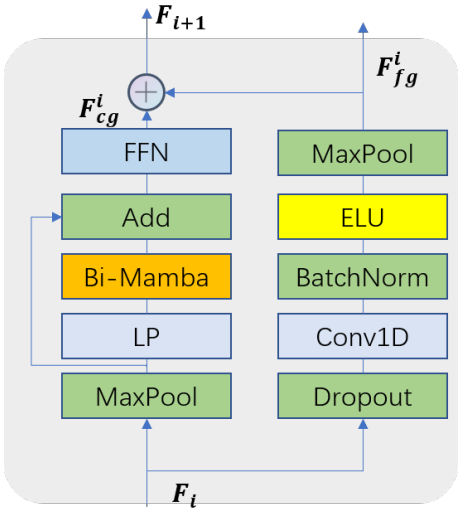

Figure 3: The structure of the Hierarchical Bi-Mamba (HBM)

Algorithm 1 Implementation of Bi-Mamba Encoder

Input: Patch-wise token sequence $E^l_x : (B,W,D)$
Output: Output features $E^{l+1}_x : (B,W,D)$

for dir in $(forward, backward)$ **do**
  if dir = backward **then**
    $E^l_{x,dir} = Flip\left(E^l_x\right)$;
  else
    $E^l_{x,dir} = E^l_x$;
  end if
  $\hat{E}^l_{y,dir} = Mamba\left(E^l_{x,dir}\right)$;
  $\hat{E}^l_{y,dir} = Add\&Norm\left(\hat{E}^l_{y,dir}, E^l_{x,dir}\right)$;
end for
$\hat{E}^{l+1}_y = E^l_{y,forward} + E^l_{y,backward}$
$\hat{E}^{l+1}_x = FeedForward\left(\hat{E}^{l+1}_y\right)$
$E^{l+1}_x = Add\&Norm\left(\hat{E}^{l+1}_y, \hat{E}^{l+1}_x\right)$
return $E^{l+1}_{x,dir}$.

Hierarchical Bi-Mamba (HBM)

Utilizing the encoded EEG features, our goal is to capture both the coarse and fine-grained temporal dynamics within EEG data by cascading HBM blocks. The architecture of an HBM block is depicted in Figure 3. An HBM comprises a parallel Bi-Mamba based branch designed to learn the correlations among the input tokens, and a CNN-based branch intended to extract fine-grained EEG features.

We define $F_i \in R^{k\times l^i_i}$ as the input of the i-th HBM block. To capture the coarse-grained temporal dynamics of EEG signals, we treat the output of each CNN kernel as one token.

Before the LP layer, a max pooling layer with a pooling size of 2 and a step of 2 is added to reduce the feature dimension of F_i. The encoding process $\Phi(\cdot)$ can be formularized by

$$E^l_x = \Phi\left(F_i\right) = MaxPool\left(F_i\right) \tag{6}$$

As shown in Figure 2. Bi-Mamba is then utilized to extract the correlations among the different views of the coarsegrained temporal embeddings. A Bi-Mamba encoder takes $E_x^l \in R^{B\times W\times D}$ as input, The detailed implementation and structure are shown in Alg. 1.

$$F_{cg}^i = FFN\left(LN\left(BiMamba\left(E_x^l\right) + MaxPool\left(F_i\right)\right)\right) \quad (7)$$

Additionally, we introduce the FTL (Ding et al., 2024) module designed to capture the short-period temporal dynamics of EEG signals. In order to extract fine-grained temporal features, we select a 1-D CNN, as its small kernel slides along the temporal dimension incrementally, facilitating the acquisition of short-period patterns. Subsequent to a dropout layer, the learned representations are passed through a 1-D CNN layer, succeeded by a batch normalization layer, an ELU activation function, and a max-pooling layer. The fine-grained temporal representations, represented as F_{fg}^i, can be computed in the following manner:

$$F_{fg}^i = MaxPool\left(ELU\left(BN\left(CNN\left(DP\left(F_i\right)\right)\right)\right)\right) \quad (8)$$

After learning the coarse- and fine-grained temporal representations, a sum fusion is added to get the final output of the HBM layer:

$$F_{i+1} = F_{cg}^i + F_{fg}^i \quad (9)$$

For the F_{fg}^i, it is also used for Dense Temporal information Fusion (DTF).

Dense Temporal information Fusion (DTF)

Leveraging the fine-grained temporal features acquired from various HBM and CNN encoder layers, we employ densely connected Dense Temporal (DT) to further extract multi-level temporal information from these layers. The goal is to distill discriminative information from a frequency standpoint and to reduce the size of the bypassed multi-level representations. Inspired by neural engineering, where the power features of EEG signals across different frequency bands are extensively utilized for analyzing brain activity (Jung et al., 1997; Schirrmeister et al., 2017). Inspired by EEG-Deformer (Ding et al., 2024), we introduce a Dense Temporal (DT) power layer for information purification and to encode frequency information within EEG signals. The power of the learned 1-D hidden representations , $I_i \in R^k$, can be calculated by

$$\begin{aligned} I_i &= DT\left(F_{fg}^i\right) \\ I_i &= log\left(\frac{1}{l_t^i}\Sigma\left(f_{fg}^{i,j}\right)^2\right), f_{fg}^{i,j} \in F_{fg}^i \end{aligned} \quad (10)$$

The refined information from all HBM layers is combined with the flattened output of the final HBM layer to produce the ultimate hidden embedding. A linear layer is subsequently employed as a classifier to map the hidden embedding onto the class labels. Let n represent the total number of HBM layers; this process can be expressed as:

$$out = Concat\left(F_n, I_0, I_1 ... I_n\right)W + b \quad (11)$$

where *Concat* is the concatenation operation, W and b are the trainable weights and biases.

Experiments

Datasets and settings

We evaluate our model on the SEED (Zheng & Lu, 2015) and SEED-V (Zhao et al., 2019) datasets, both with 62 electrode channels and video-induced emotions. SEED has 15 subjects, 3 sessions, 15 trials per session, with positive, negative and neutral emotions. SEED-V has 16 subjects, same structure, but with happy, disgust, sad, neutral and fear emotions. EEG signals are segmented into 4-second non-overlapping windows. Experiments are subject-dependent and follow the train-test splits of (Zheng & Lu, 2015) and (Zhao et al., 2019).

Training uses cross-entropy loss, Adam optimizer (initial learning rate 1e-3, weight decay 1e-5), cosine annealing for dynamic learning rate, and dropout rate 0.5 to prevent overfitting. Batch size is 64, trained for 200 epochs, with the highest validation accuracy model used for testing. CNN kernel lengths are $0.1 \times f_s$, where f_s is the EEG sampling rate.

Baselines

The comparison baselines are as follows: 1. SVM (Y. Li et al., 2020): A machine learning method using discriminant entropy features with an SVM classifier. 2. BDAE (Zhao et al., 2019): A bimodal deep autoencoder for extracting high-level emotion representations. 3. R2G-STNN (Y. Li et al., 2019): Extracts spatial relationships within/between brain regions and dynamic temporal info. 4. BiHDM (X. Li et al., 2022): Explores left/right hemisphere feature differences and captures temporal info in two directions via RNNs. 5. RGNN (Zhong et al., 2020): A regularized GNN examining EEG channel topology using two regularizers. 6. 4D-CRNN (Shen et al., 2020): Combines CNN and RNN to extract spatial, spectral, and temporal EEG features. 7. MD-AGCN (R. Li et al., 2021): An adaptive GCN fusing frequency and temporal domain features. 8. PGCN (Jin et al., 2023): A graph convolution model integrating local, mesoscopic, and global features at various scales. 9. CSET-CCA (Pan & Bai, 2024): EEG emotion recognition via convolutional transformer with class confusion-aware attention.

Results

Classification Results

We compare the proposed Emotion-Mamba model with baselines on the SEED and SEED-V datasets. Table 1 shows the accuracy (ACC) and standard deviation (STD) of these models on the SEED and SEED-V datasets. Compared with that of the baseline models, the ACC of our model is further improved. Emotion-Mamba achieves state-of-theart performance. On the SEED dataset, an ACC of 95.33% and an STD of 4.25% are achieved.

The Emotion-Mamba simultaneously considers multi-level temporal information, the model refines the fine-grained temporal representations obtained from all HBM layers and the

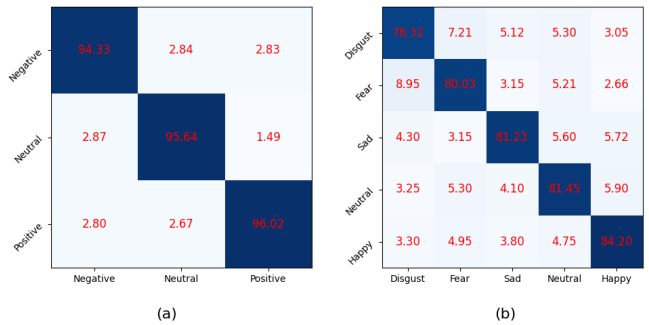

Figure 4: Confusion matrices of the Emotion-Mamba. (a) Confusion matrix on the SEED. (b) Confusion matrix on the SEED-V.

Table 1: Comparison of classification performance on the SEED and SEED-V datasets

Models	SEED ACC ± STD (%)	SEED-V ACC ± STD (%)
SVM	83.99 ± 9.72	69.50 ± 10.28
BADE	\	79.70 ± 4.76
R2G-STNN	93.38 ± 5.96	\
BiHDM	93.12 ± 6.06	\
RGNN	94.24 ± 5.95	\
4D-CRNN	94.74 ± 2.32	\
MD-AGCN	94.81 ± 4.52	80.77 ± 6.61
PGCN	\	81.29 ± 10.57
CSET-CCA	94.88 ± 4.35	81.46 ± 8.42
Emotion-Mamba	**95.33 ± 4.25**	**82.21 ± 6.32**

CNN encoder which enables our model to adequately capture valuable information in EEG signals, Our model still achieves the best result on the SEED-V dataset, with an ACC of 82.21% and an STD of 6.32%.

Figure 4 depicts the confusion matrices of the model on these two datasets. On the SEED dataset, our model demonstrates stronger performance in recognizing positive and neutral emotions, achieving accuracies of 96.02% and 95.64% respectively, compared to 94.33% for negative emotion recognition.

Regarding the SEED-V dataset, although our model has a relatively lower recognition accuracy of 78.32% for the disgust emotion, this result is still remarkable. Given that disgust is a rather confounding emotion class, and only 7.21% of samples are misclassified as fear and 3.05% as happy. Moreover, the recognition accuracies for the other emotion classes remain relatively high.

Ablation Study

We follow the previous work (Ding et al., 2024), retaining Dense Temporal information (DT) in four stages. We conduct ablation experiments from the following aspects: whether to use Dense Temporal Fusion (DTF), whether to use plain Mamba replace Bi-Mamba, and whether to use DT in the CNN Encoder for time information extraction. We conduct ablation experiments on the SEED dataset, and the experimental results are shown in Table 2. When the Dense Temporal Fusion (DTF) module is not used, the accuracy (ACC) decreases by 5.06%. Integrating dense temporal information can significantly enhance the performance of the model. When we do not adopt the temporal information from the CNN Encoder, the accuracy will drop by 0.6%. It can be seen that incorporating more temporal information is crucial for the model. When we use plain Mamba instead of Bi-Mamba, the ACC decreases by 2.12%. Therefore, the bidirectional Mamba module can fuse multi-stage information more effectively.

Table 2: Ablation study on the SEED dataset

	ACC ± STD (%)	Variations(%)
w/o DTF	90.27± 6.43	-5.06
w/o CNN DT	94.73± 4.12	-0.6
w/o Bi-Mamba	93.21 ± 5.11	-2.12
Emotion-Mamba	**95.33 ± 4.25**	0

Conclusion

In this paper, we present Emotion-Mamba, a novel model for emotion recognition. Initially, the model processes data through a CNN encoder. Subsequently, a sequence of Hierarchical Bi-Mamba (HBM) modules come into play. These modules are meticulously designed to capture both coarse-grained and fine-grained temporal dynamics in EEG signals. To efficiently leverage multi-level temporal information, the model refines the fine-grained temporal representations obtained from the CNN encoder and all HBM layers. These refined signals are then densely connected to the final embeddings. Subsequently, they enter the classifier for emotion classification. Experimental results clearly demonstrate that our model achieves state-of-the-art classification performance on benchmark datasets. The results of ablation experiments further confirm the effectiveness of each module. In future work, we will further optimize the model and enhance its generalization ability, aiming to increase its practical value in real-world applications.

Acknowledgement

This work is supported by the National Key Research and Development Project of China No. 2023YFC3107100, NSFC No. 62172203, the Collaborative Innovation Center of Novel Software Technology and Industrialization. (*Liang Wang is the corresponding author. wl@nju.edu.cn)

References

Burle, B., Spieser, L., Roger, C., Casini, L., Hasbroucq, T., & Vidal, F. (2015). Spatial and temporal resolutions of eeg: Is it really black and white? a scalp current density view. *International Journal of Psychophysiology*.

Ding, Y., & Guan, C. (2023). Gign: Learning graph-in-graph representations of eeg signals for continuous emotion recognition.

Ding, Y., Robinson, N., Tong, C., Zeng, Q., & Guan, C. (2023). Lggnet: Learning from local-global-graph representations for brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems*.

Ding, Y., Robinson, N., Zhang, S., Zeng, Q., & Guan, C. (2023). Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*.

Ding, Y., Li, Y., Sun, H., Liu, R., Tong, C., Liu, C., Zhou, X., & Guan, C. (2024). Eeg-deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*.

Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors*.

Feng, L., Cheng, C., Zhao, M., Deng, H., & Zhang, Y. (2022). Eeg-based emotion recognition using spatial-temporal graph convolutional lstm with attention mechanism. *IEEE Journal of Biomedical and Health Informatics*.

Foong, R., Ang, K. K., Quek, C., Guan, C., Phua, K. S., Kuah, C. W. K., Deshmukh, V. A., Yam, L. H. L., Rajeswaran, D. K., & Tang, N. (2019). Assessment of the efficacy of eeg-based mi-bci with visual feedback and eeg correlates of mental fatigue for upper-limb stroke rehabilitation. *IEEE Transactions on Biomedical Engineering*.

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A., Goel, K., Gupta, A., & Ré, C. (2022). On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, *35*, 35971–35983.

Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., & Ré, C. (2021). Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, *34*, 572–585.

Gupta, A., Gu, A., & Berant, J. (2022). Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, *35*, 22982–22994.

He, J., Zhao, L., Yang, H., Zhang, M., & Li, W. (2019). Hsibert: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Transactions on Geoscience and Remote Sensing*.

Jia, Z., Liang, H., Liu, Y., Wang, H., & Jiang, T. (2024). Distillsleepnet: Heterogeneous multi-level knowledge distillation via teacher assistant for sleep staging. *IEEE Transactions on Big Data*.

Jia, Z., Wang, H., Liu, Y., & Jiang, T. (2024). Mutual distillation extracting spatial-temporal knowledge for lightweight multi-channel sleep stage classification.

Jin, M., Zhu, E., Du, C., He, H., & Li, J. (2023). Pgcn: Pyramidal graph convolutional network for eeg emotion recognition. *arXiv preprint arXiv:2302.02520*.

Jung, T.-P., Makeig, S., Stensmo, M., & Sejnowski, T. (1997). Estimating alertness from the eeg power spectrum. *IEEE Transactions on Biomedical Engineering*.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*.

Lee, Y.-E., & Lee, S.-H. (2022). Eeg-transformer: Self-attention from transformer architecture for decoding eeg of imagined speech.

Li, R., Wang, Y., & Lu, B.-L. (2021). A multi-domain adaptive graph convolutional network for eeg-based emotion recognition.

Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., et al. (2022). Eeg based emotion recognition: A tutorial and review. *ACM Computing Surveys*.

Li, Y., Wang, L., Zheng, W., Zong, Y., Qi, L., Cui, Z., et al. (2020). A novel bi-hemispheric discrepancy model for eeg emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*.

Li, Y., Zheng, W., Wang, L., Zong, Y., & Cui, Z. (2019). From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., & Liu, Y. (2024). Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.

Lotte, F., & Guan, C. (2010). Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering*.

Malfliet, A., Coppieters, I., Van Wilgen, P., Kregel, J., De Pauw, R., Dolphens, M., & Ickmans, K. (2017). Brain changes associated with cognitive and emotional factors in chronic pain: A systematic review. *European Journal of Pain*.

Mane, R., Robinson, N., Vinod, A. P., Lee, S.-W., & Guan, C. (2020). A multiview cnn with novel variance layer for motor imagery brain computer interface.

Mitchell, J. (2021). Affective shifts: Mood, emotion and well-being. *Synthese*.

Ozdemir, M. A., Degirmenci, M., Izci, E., & Akan, A. (2021). Eeg-based emotion recognition with deep convolutional neural networks. *Biomedical Engineering/Biomedizinische Technik*.

Pan, J., & Bai, C. (2024). Eeg-based emotion recognition via convolutional transformer with class confusion-aware attention. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Riberto, M., Paz, R., Pobric, G., & Talmi, D. (2022). The neural representations of emotional experiences are more similar than those of neutral experiences. *Journal of Neuroscience*.

Ruan, J., & Xiang, S. (2024). Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*.

Shen, F., Dai, G., Lin, G., Zhang, J., Kong, W., & Zeng, H. (2020). Eeg-based emotion recognition using 4d convolutional recurrent neural network. *Cognitive Neurodynamics*.

Song, T., Zheng, W., Song, P., & Cui, Z. (2020). Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*.

Song, Y., Zheng, Q., Liu, B., & Gao, X. (2022). Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Song, Y., Zheng, Q., Liu, B., & Gao, X. (2023). Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Tao, W., Li, C., Song, R., Cheng, J., Liu, Y., Wan, F., & Chen, X. (2020). Eeg-based emotion recognition via channelwise attention and self attention. *IEEE Transactions on Affective Computing*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need.

Xefteris, V., Tsanousa, A., Georgakopoulou, N., Diplaris, S., Vrochidis, S., & Kompatsiaris, I. (2022). Graph theoretical analysis of eeg functional connectivity patterns and fusion with physiological signals for emotion recognition. *Sensors*.

Xie, J., Zhang, J., Sun, J., Ma, Z., Qin, L., Li, G., Zhou, H., & Zhan, Y. (2022). A transformer-based approach combining deep learning network and spatial-temporal information for raw eeg classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Zhao, L., Li, R., Zheng, W., & Lu, B. (2019). Classification of five emotions from eeg and eye movement signals: Complementary representation properties.

Zheng, W., & Lu, B. (2015). Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*.

Zhong, P., Wang, D., & Miao, C. (2020). Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*.

Zotev, V., Mayeli, A., Misaki, M., & Bodurka, J. (2020). Emotion self-regulation training in major depressive disorder using simultaneous real-time fmri and eeg neurofeedback. *NeuroImage: Clinical*.