BraSTORM: A Dual-Branch Self-Supervised Framework for EEG Representation Learning via Input-Level Spatio-Temporal Decomposition

Yifan Wang[1, 2], **Der-Horng Lee**[2*], **Bruce X.B. Yu**[2*]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027 China
[2] ZJU-UIUC Institute, Zhejiang University, Haining, 314400 China
{yifan3.23, dhlee, xinboyu}@intl.zju.edu.cn

Abstract

Prevalent pre-training strategies for Brain-Computer Interfaces (BCIs) are often constrained by spatio-temporal entanglement. This critical issue arises from processing multi-channel Electroencephalography (EEG) signals as monolithic sequences, which intertwines the signal's temporal dynamics with its spatial topography and hinders the learning of robust and generalizable representations. To address this, we introduce BraSTORM, a framework that explicitly disentangles EEG data into separate temporal and spatial streams at the input level. Two streams are processed by parallel encoders trained with a composite dual-objective: a masked signal reconstruction loss captures fine-grained, intra-modal details, while a cross-modal contrastive loss enforces high-level semantic alignment. Extensive fine-tuning experiments on six benchmarks covering three major BCI downstream tasks—Emotion Recognition, Sleep Staging, and Motor Imagery—demonstrate that BraSTORM achieves state-of-the-art performance. Our findings validate that resolving spatio-temporal entanglement at the input level can be a competitive pre-training framework for the BCI field.

Introduction

Deep learning now underpins modern Brain-Computer Interface (BCI) systems (Hwang et al. 2013; Subha et al. 2010) for Electroencephalography (EEG) analysis. Models based on this paradigm excel at critical tasks, including motor imagery (MI) classification (Altaheri et al. 2023), emotion recognition (Prabowo et al. 2023), and automatic sleep staging (Phan and Mikkelsen 2022). This success stems from the capacity of architectures, such as Convolutional Neural Networks (CNNs) (Yang et al. 2018), Graph Convolutional Networks (GCNs) (Song et al. 2018), and Transformers (Song et al. 2022), to automatically learn discriminative representations from high-dimensional EEG data. Consequently, these supervised models have consistently pushed state-of-the-art benchmarks in decoding brain activity.

However, these supervised paradigms are fundamentally constrained by two critical challenges. Firstly, its heavy reliance on vast, meticulously labeled datasets is often impractical, as EEG data collection is notoriously labor-intensive and prone to subjective bias (Zhang, Zhong, and Liu 2024a; Wang, Chen, and Song 2024). Secondly, these models grapple with significant inter-subject variability, where neural patterns differ drastically across individuals and even sessions (Kostas and Rudzicz 2021). This "domain shift" phenomenon severely hampers their ability to generalize to unseen subjects, creating a critical bottleneck for developing truly robust and scalable BCI systems (Wang, Zhang, and Tang 2024). Consequently, the field is actively seeking alternative learning paradigms.

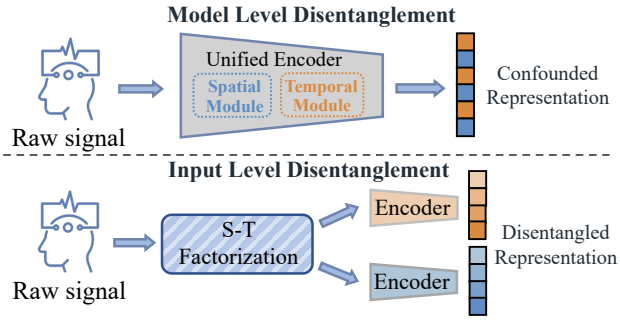

Figure 1: The paradigm differences between implicit model-level disentanglement and BraSTORM's explicit input-level factorization for EEG representation learning.

To circumvent these limitations, Self-Supervised Learning (SSL) has emerged as a transformative paradigm. By leveraging vast quantities of unlabeled data, SSL enables the pre-training of large-scale "foundation models" that learn rich, generalizable representations (Lai et al. 2025), which can then be efficiently adapted to downstream tasks with minimal labeled data. Inspired by successes in language and vision, recent efforts have focused on pre-training large Transformer-based models (Jiang, Zhao, and liang Lu 2024; Chen et al. 2024) for EEG. A dominant approach is Masked Signal Modeling (MSM), where models learn by reconstructing corrupted portions of the input, as demonstrated by pioneering works (Zhang et al. 2023; Cui et al. 2024). Despite their promise, these approaches inherently follow a paradigm of model level disentanglement (Figure 1, top). They process multi-channel EEG as a monolithic stream, entangling spatial topography with temporal dynamics and

*Corresponding authors.

compelling the model to learn from confounded signals.

Inspired by recent work that successfully disentangled EEG into time and frequency domains for robust learning (Liu et al. 2024b), we claim that resolving this spatio-temporal entanglement at the input level is a more fundamental and effective strategy. This principle motivates BraSTORM, our framework designed for explicit input level disentanglement (Figure 1, bottom). It first factorizes the signal into separate spatial and temporal streams, then employs a composite dual-objective—combining masked reconstruction and cross-modal contrastive learning—to produce representations that are robust, interpretable, and transferable. The contributions are summarized as follows:

- We identify and formalize spatio-temporal entanglement as a bottleneck in current EEG foundation models. To solve this, we propose BraSTORM, a self-supervised framework built on the principle of input-level disentanglement.
- We devise a composite pre-training strategy that combines the specialization of dual-stream masked signal modeling with the semantic alignment of cross-modal contrastive learning, creating representations that are both specific and holistic.
- We establish the superiority of our disentanglement approach through extensive experiments on six benchmarks across three major BCI paradigms, where BraSTORM consistently achieves state-of-the-art performance and demonstrates superior generalization.

Related Work

Supervised Learning & Generalization

Deep learning has become a standard for EEG signal decoding, with specialized supervised models achieving strong task-specific results. For instance, architectures like EEG Conformer (Song et al. 2022) and TSception (Ding et al. 2022) excel at decoding complex neural patterns. However, their success is predicated on large, labeled datasets, and their performance often degrades when faced with inter-subject variability, a major challenge in BCI (Altaheri et al. 2023). This critical limitation of poor generalization has catalyzed a field-wide shift towards the pre-train, fine-tune paradigm enabled by SSL.

Self-Supervised Learning Paradigms

The SSL landscape is shaped by two complementary paradigms: (1) Contrastive learning (e.g., SimCLR (Chen et al. 2020)) learns invariant representations by maximizing agreement between augmented views of an input. This has been adapted to EEG by methods like SeqCLR (Mohsenvand, Izadi, and Maes 2020) and JCFA (Liu et al. 2024b) to learn global semantic features. However, its focus on invariance risks discarding fine-grained local patterns. (2) masked modeling (e.g., BERT (Devlin et al. 2019), MAE (He et al. 2022)) excels at learning rich, contextual local structures by forcing an encoder to reconstruct missing input portions, as seen in MAEEG (Chien et al. 2022). Yet, it may not effectively capture high-level abstract semantics. This reveals a trade-off, highlighting that neither approach is sufficient alone. A robust representation necessitates a composite framework that integrates the local detail from reconstruction with the global alignment from contrastive learning.

Spatio-Temporal Entanglement

Recent SSL models for EEG, such as BIOT (Yang, Westover, and Sun 2023) and LaBraM (Jiang, Zhao, and liang Lu 2024), typically treat EEG as a monolithic sequence. This creates spatio-temporal entanglement, where learned features conflate temporal dynamics with spatial topography, leading to fragile representations. Recognizing this, later works like EEGPT (Wang et al. 2024) and CBraMod (Wang et al. 2025) use internal architectural mechanisms to separate dependencies. However, since they still operate on a unified input stream, they only attempt to disentangle information within the model, not at the signal's source. Other graph-based methods (Wei et al. 2024) distill topological priors rather than disentangling the signal's intrinsic properties.

Thus, a critical, unaddressed challenge remains: learning from representations that are explicitly and fundamentally decoupled at the input stage. Our work is the first to address this gap, proposing a dual-branch framework that tackles entanglement at its root.

Methodology

Our proposed method, BraSTORM, operates by factorizing EEG signals into distinct spatial and temporal streams. This data-level decoupling directly addresses the inherent spatio-temporal entanglement problem. As illustrated in Figure 2, our framework is composed of three core components (ordered from left to right): (1) a dual-stream spatio-temporal factorization module, (2) a composite pre-training stage driven by a dual objective, and (3) a fine-tuning stage for downstream adaptation.

Dual-Stream Spatio-Temporal Factorization

Temporal Stream Input The input EEG sample represented as a matrix $X \in \mathbb{R}^{C\times T}$ (where C is the number of channels and T is the number of time points) is partitioned along the temporal axis into N_p non-overlapping patches, yielding a sequence of patches $x^T = \{p_1, p_2, \ldots, p_{N_p}\}$. Each patch $p_i \in \mathbb{R}^{C\times L}$, with patch length $L = T/N_p$, represents a local time window of multi-channel brain activity. This representation, x^T, is designed to preserve the high-frequency components and dynamic relationships within the signal, providing an input for learning temporal patterns.

Spatial Stream Input To isolate spatial topographies, we transform the raw signal X into a compact, feature-rich "spatio-spectral cube". This process involves two steps. First, for each channel c and frequency band $b \in \{\theta, \alpha, \beta, \gamma\}$, we compute the Differential Entropy (DE) (Zheng and Lu 2015). Assuming the band-filtered signal vector for a specific channel and band, denoted as $x_{c,b}$, follows a Gaussian distribution with variance $\sigma^2_{c,b}$, its DE is

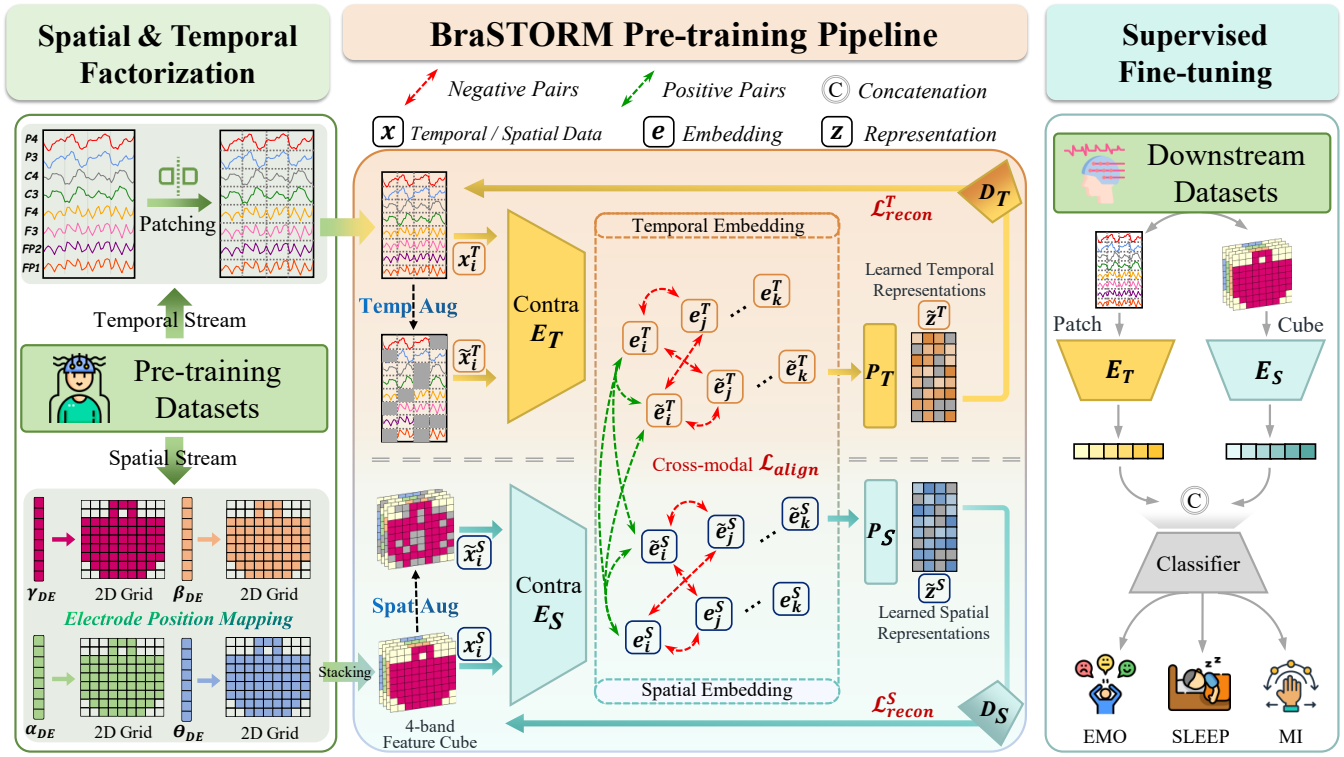

Figure 2: The overall framework of BraSTORM. **(1) Input Factorization:** Raw EEG signals are decomposed into a temporal sequence of patches and a corresponding spatio-spectral cube. **(2) Self-supervised Pre-training:** Two separate encoders, E_T and E_S, are trained with a dual objective: cross-modal alignment ($\mathcal{L}_{\text{align}}$) and intra-modal reconstruction ($\mathcal{L}_{\text{recon}}$). **(3) Supervised Fine-tuning:** The pre-trained encoders are used as feature extractors for downstream classification tasks.

defined as:

$$\mathrm{h}(x_{c,b}) = \frac{1}{2}\log(2\pi e\sigma^2_{c,b}) \tag{1}$$

This operation, performed for all C channels, yields a feature matrix $F_{\text{DE}} \in \mathbb{R}^{C\times 4}$, summarizing the spectral power distribution across all channels for the four frequency bands. Second, to explicitly encode spatial relationships, we project the DE features of each band onto a $2D$ grid that mimics the sensor layout, creating four spatial maps: G_θ, G_α, G_β and G_γ. These maps are then stacked to form the final spatial stream input:

$$x^S = \text{Stack}(G_\theta, G_\alpha, G_\beta, G_\gamma) \in \mathbb{R}^{4\times H\times W} \tag{2}$$

This representation provides a rich, multi-band snapshot of the brain's topographical activity, primed for spatial feature extraction.

Composite Pre-training Objective

The pre-training of BraSTORM is driven by a unified objective function designed to learn representations that are both rich in domain-specific detail and semantically aligned across modalities. For a mini-batch of size N, denoted as $\{X_i\}_{i=1}^N$, drawn from the training dataset $\mathcal{D}$ we first generate their factorized views, which we denote as the "original" views: a temporal patch sequence x_i^T and a spatial feature cube x_i^S. We then create augmented (masked) views, $\tilde{x}_i^T$ and $\tilde{x}_i^S$. For the temporal stream $x^T = \{p_1, \ldots, p_{N_p}\}$, we randomly select a subset of patch indices with a masking ratio ρ_{t} and replace the corresponding patches with a full-zero token $p_{\text{mask}} \in \mathbb{R}^{C\times L}$. A similar patch-level masking strategy with ratio ρ_{s} is applied to the spatial stream input x^S before it is fed to its encoder. These original and augmented views are then processed by two parallel branches, each with an encoder specialized for its respective modality.

Temporal Branch. The temporal encoder, denoted as E_T, is an Attention-based Temporal Convolutional Network (Altaheri, Muhammad, and Alsulaiman 2022). It models the inter-relationships among the sequence of temporal patches to capture complex dependencies. **Spatial Branch.** The spatial encoder, E_S, is a Vision Transformer (ViT) (Dosovitskiy et al. 2020). It processes the $4 \times H \times W$ spatio-spectral feature cube x^S by dividing it into a sequence of non-overlapping patches and applying patch-based attention to learn the topographical relationships.

Cross-Modal Alignment To learn semantically aligned representations, we introduce a cross-modal alignment loss, $\mathcal{L}_{\text{align}}$. This objective synchronizes the high-level information captured by the temporal and spatial encoders. For each sample X_i in a mini-batch, we obtain four embeddings by passing its four views through their respective encoders:

$$\begin{aligned} \mathbf{e}_i^T &= \mathrm{E}_T(\mathbf{x}_i^T), \quad \tilde{\mathbf{e}}_i^T = \mathrm{E}_T(\tilde{\mathbf{x}}_i^T), \\ \mathbf{e}_i^S &= \mathrm{E}_S(\mathbf{x}_i^S), \quad \tilde{\mathbf{e}}_i^S = \mathrm{E}_S(\tilde{\mathbf{x}}_i^S) \end{aligned} \tag{3}$$

The core principle is that all four embeddings derived from the same source sample should be attracted to each other while being repelled from the embeddings of other samples. We adopt a multi-positive contrastive loss. For any given embedding (anchor) e_a, its three counterparts from the same source sample constitute its positive set $\mathcal{P}(a)$, while all other embeddings in the mini-batch form the negative set $\mathcal{N}(a)$. The loss for anchor e_a is defined as:

$$\mathcal{L}(\mathbf{e}_a) = -\frac{1}{|\mathcal{P}(a)|} \sum_{\mathbf{e}_p \in \mathcal{P}(a)} \log \left(\frac{\exp(\mathrm{sim}(\mathbf{e}_a, \mathbf{e}_p)/\tau)}{S_a} \right) \tag{4}$$

$$\text{where} \quad S_a = \sum_{\mathbf{e}_k \in \mathcal{P}(a) \cup \mathcal{N}(a)} \exp(\mathrm{sim}(\mathbf{e}_a, \mathbf{e}_k)/\tau) \tag{5}$$

where $\mathrm{sim}(\cdot, \cdot)$ is the cosine similarity and τ is a temperature hyperparameter. The total alignment loss is the average loss over all $4N$ embeddings in the mini-batch:

$$\mathcal{L}_{\text{align}} = \frac{1}{4N} \sum_{i=1}^{N} \sum_{\mathbf{e} \in \left\{ \mathbf{e}_i^T, \tilde{\mathbf{e}}_i^T, \mathbf{e}_i^S, \tilde{\mathbf{e}}_i^S \right\}} \mathcal{L}(\mathbf{e}) \tag{6}$$

Intra-Modal Reconstruction To compel the primary encoders to capture rich, fine-grained structural details, we introduce a complementary masked reconstruction objective, $\mathcal{L}_{\text{recon}}$. This task operates on the final representations produced by two lightweight MLP projectors, P_T and P_S. These projectors map the high-dimensional embeddings of the augmented views ($\tilde{\mathbf{e}}_i^T$, $\tilde{\mathbf{e}}_i^S$) to a latent space. The resulting projected representations are:

$$\tilde{\mathbf{z}}_i^T = P_T(\tilde{\mathbf{e}}_i^T) \quad , \quad \tilde{\mathbf{z}}_i^S = P_S(\tilde{\mathbf{e}}_i^S) \tag{7}$$

These are then fed into two lightweight Transformer decoders, D_T and D_S, which are tasked with reconstructing the original, unmasked inputs. The reconstruction loss is the sum of the Mean Squared Error (MSE) for each stream:

$$\mathcal{L}_{\text{recon}}^T = \frac{1}{N} \sum_{i=1}^{N} \left\| D_T(\tilde{\mathbf{z}}_i^T) - \mathbf{x}_i^T \right\|_2^2 \tag{8}$$

$$\mathcal{L}_{\text{recon}}^S = \frac{1}{N} \sum_{i=1}^{N} \left\| D_S(\tilde{\mathbf{z}}_i^S) - \mathbf{x}_i^S \right\|_2^2 \tag{9}$$

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{recon}}^T + \mathcal{L}_{\text{recon}}^S \tag{10}$$

This objective compels the entire network, from encoder to projector, to retain high-fidelity information. The final objective for BraSTORM is a weighted combination of these two losses, optimized end-to-end:

$$\mathcal{L}_{\text{BraSTORM}} = \mathcal{L}_{\text{align}} + \lambda_{\text{rec}} \mathcal{L}_{\text{recon}} \tag{11}$$

where scalar λ_{rec} is balances two objectives.

Phase	Dataset	Type	# Channels	# Samples	Duration
Pre-training	SEED	EMO	62	152,730	1s
	SEED-IV	EMO	62	151,845	1s
	SEED-V	EMO	62	115,001	1s
	M3CV	MULTI	64	231,404	1s
	TSU	SSVEP	64	50,400	1s
Fine-tuning	FACED	EMO	30	10,332	10s
	DEAP	EMO	32	15,360	4s
	ISRUC	SLEEP	6	113,283	30s
	HMC	SLEEP	4	106,954	30s
	BCIC-IV 2a	MI	22	5,184	7s
	SPMI	MI	30	4,000	2s

Table 1: Overview of datasets utilized for pre-training and downstream fine-tuning.

Adaptation for Downstream Tasks

To validate the learned representations, we adapt the pre-trained encoders for downstream tasks using two standard strategies: linear probing (Wang et al. 2024) and full fine-tuning. In both settings, for any given input, temporal and spatial feature vectors are extracted by their respective encoders (E_T, E_S) and then concatenated. This fused spatio-temporal representation is fed into a task-specific classification head. For linear probing, the pre-trained encoders are frozen, and only the new linear classifier is trained, measuring the raw quality of the fixed features. For full fine-tuning, the entire architecture, including the encoders, is trained end-to-end, typically with a lower learning rate for the backbones to retain pre-trained knowledge. This dual evaluation allows a comprehensive assessment of the representations.

Experiments

To validate the effectiveness and generalizability of the proposed BraSTORM framework, extensive experiments are conducted across a large-scale self-supervised pre-training phase and a subsequent evaluation on multiple downstream BCI tasks.

Experiment Setup

Pre-training Setup The pre-training corpus is constructed by aggregating five public datasets: SEED (Zheng and Lu 2015), SEED-IV (Zheng et al. 2018), SEED-V (Liu et al. 2021), M3CV (Huang et al. 2022), and TSU (Wang et al. 2016). This corpus totals over $701,000$ one-second EEG samples, covering paradigms like emotion recognition (EMO), steady-state visual evoked potential (SSVEP), and multi-task recordings. An overview of each dataset is presented in Table 1.

A standardized preprocessing pipeline was applied across all datasets. All recordings were re-referenced to a Common Average Reference (CAR), band-pass filtered between 0.5 Hz and 45 Hz. The signals were then resampled to 200 Hz and segmented into non-overlapping 1-second windows. For the spatial stream, EEG channels were mapped onto a 9×9 grid based on their standard topographical locations. Finally, all samples underwent channel-wise z-score normalization.

EMO	FACED, 9-class			DEAP, 4-class		
Methods	**Balanced Acc.**	**Cohen's Kappa**	**Weighted-F1**	**Balanced Acc.**	**Cohen's Kappa**	**Weighted-F1**
EEGNet	43.24 ± 2.17	36.74 ± 2.40	42.39 ± 1.79	47.47 ± 1.98	21.05 ± 0.51	42.02 ± 2.75
TSCeption	45.53 ± 1.95	36.39 ± 1.99	31.92 ± 1.70	45.76 ± 1.53	21.01 ± 0.87	42.24 ± 1.86
EEGConformer	56.41 ± 0.91	47.63 ± 1.18	51.46 ± 1.17	44.20 ± 0.91	20.87 ± 0.81	47.10 ± 0.92
CSPNet	46.93 ± 1.01	39.98 ± 1.75	46.34 ± 0.93	48.03 ± 1.12	22.39 ± 0.96	48.29 ± 1.35
BENDR	55.15 ± 1.02	52.58 ± 1.24	54.72 ± 1.13	50.14 ± 1.28	24.51 ± 0.95	50.88 ± 1.52
BIOT	51.18 ± 1.18	44.76 ± 2.54	51.36 ± 1.12	49.55 ± 1.41	23.87 ± 1.10	49.89 ± 1.63
LaBraM-Base	57.23 ± 0.98	52.99 ± 1.05	56.84 ± 1.01	52.16 ± 1.19	27.82 ± 0.91	52.50 ± 1.27
EEGPT	57.54 ± 1.21	52.81 ± 1.30	56.90 ± 1.25	51.57 ± 1.33	26.93 ± 1.04	51.96 ± 1.48
Our BraSTORM	**59.12 ± 0.85**	**53.58 ± 0.99**	**58.39 ± 0.88**	**55.82 ± 1.05**	**31.98 ± 0.89**	**56.01 ± 1.14**
SLEEP	ISRUC, 5-class			HMC, 5-class		
EEGNet	61.40 ± 1.51	48.85 ± 1.82	57.00 ± 1.65	57.44 ± 2.11	42.30 ± 2.53	55.26 ± 1.98
TSCeption	64.04 ± 1.35	57.73 ± 1.40	66.83 ± 1.29	52.36 ± 1.87	39.30 ± 2.01	50.41 ± 1.75
EEGConformer	65.49 ± 1.12	56.70 ± 1.21	66.82 ± 1.18	58.16 ± 1.55	53.08 ± 1.49	58.20 ± 1.62
CSPNet	63.24 ± 1.42	55.03 ± 1.55	64.53 ± 1.38	53.44 ± 1.78	40.14 ± 1.92	51.29 ± 1.66
BENDR	68.22 ± 1.05	59.89 ± 1.15	67.95 ± 1.09	63.10 ± 1.33	58.52 ± 1.41	62.88 ± 1.25
BIOT	68.95 ± 1.24	60.83 ± 1.31	68.50 ± 1.28	64.01 ± 1.45	59.76 ± 1.50	63.95 ± 1.39
LaBraM-Base	71.38 ± 0.95	63.50 ± 1.02	70.91 ± 0.98	67.29 ± 1.18	62.44 ± 1.23	66.99 ± 1.11
EEGPT	70.51 ± 1.01	62.17 ± 1.19	69.88 ± 1.06	66.83 ± 1.21	61.90 ± 1.35	65.74 ± 1.28
Our BraSTORM	**71.86 ± 0.89**	**66.92 ± 0.97**	**73.15 ± 0.91**	**70.59 ± 1.02**	**64.14 ± 1.15**	**69.35 ± 0.98**
MI	BCIC-IV 2a, 4-class			SPMI, 2-class (Balanced Acc, AUC-PR, AUROC)		
EEGNet	42.24 ± 2.05	22.98 ± 2.31	42.26 ± 2.10	52.50 ± 1.85	54.75 ± 1.55	61.27 ± 1.60
TSCeption	31.38 ± 1.91	8.52 ± 2.45	31.35 ± 1.88	54.75 ± 1.62	53.84 ± 1.49	60.49 ± 1.53
EEGConformer	48.59 ± 1.43	31.94 ± 1.67	49.09 ± 1.39	53.62 ± 1.35	52.30 ± 1.28	59.71 ± 1.41
CSPNet	36.03 ± 1.82	14.71 ± 2.13	35.42 ± 1.77	49.75 ± 2.01	49.77 ± 1.98	57.03 ± 1.95
BENDR	49.12 ± 1.35	32.50 ± 1.44	49.33 ± 1.31	61.05 ± 1.28	56.89 ± 1.35	68.14 ± 1.22
BIOT	49.88 ± 1.41	33.15 ± 1.52	49.95 ± 1.38	62.34 ± 1.33	57.55 ± 1.41	68.01 ± 1.30
LaBraM-Base	50.95 ± 1.15	35.21 ± 1.29	**50.80** ± 1.11	65.11 ± 1.08	59.83 ± 1.12	71.54 ± 1.01
EEGPT	44.24 ± 1.23	28.58 ± 1.38	44.19 ± 1.25	64.29 ± 1.15	59.01 ± 1.24	71.33 ± 1.19
Our BraSTORM	**51.59 ± 1.21**	**37.17 ± 1.33**	50.59 ± 1.28	**68.07 ± 0.95**	**61.19 ± 1.08**	**73.29 ± 0.91**

Table 2: Comprehensive experimental results for emotion recognition, sleep staging, and motor imagery classification. All metrics are presented as mean ± std (%). The best results are in bold, and the best baseline results are underlined.

BraSTORM was pre-trained for 200 epochs with a batch size of 128. The AdamW optimizer was used with a learning rate of 5e-4, a weight decay of 1e-5, and a cosine annealing scheduler. The reconstruction loss weight λ_{rec} was set to 0.6, and the contrastive temperature τ was set to 0.07. The masking ratios ρ_{t} and ρ_{s} were set to 0.6 and 0.5, respectively. All experiments were implemented in Python 3.12.1, using PyTorch-Lightning 2.4.1 and Torcheeg 1.1.3 (Zhang, Zhong, and Liu 2024b). The training and evaluation were conducted on a server equipped with 10 NVIDIA RTX 4090 GPUs, each with 24 GB of VRAM.

Fine-tuning Setup BraSTORM is evaluated on six benchmark datasets across three BCI applications: (1) Emotion Recognition (EMO): FACED (Chen et al. 2023) and DEAP (Koelstra et al. 2011); (2) Sleep Staging (SLEEP): ISRUC (Khalighi et al. 2016) and HMC (Alvarez-Estevez and Rijsman 2021); and (3) Motor Imagery (MI) Classification: BCIC-IV 2a (Brunner et al. 2008) and SPMI (Liu et al. 2024a). Further details are summarized in Table 1.

The evaluation follows a strict subject-independent paradigm. For each dataset, subjects were randomly split into training (80%), validation (10%), and test (10%) sets. We assess the quality of the pre-trained representations using two standard protocols: linear probing and full fine-tuning. For this fine-tuning stage, the AdamW optimizer was used with a learning rate of 1e-4 and a batch size of 64. Training was run for a maximum of 100 epochs with an early stopping strategy (patience of 10 epochs).

Main Results

Baselines and Evaluation Metrics To assess the effectiveness of BraSTORM, its performance is benchmarked against two categories of models: established supervised methods (i.e., EEGNet (Lawhern et al. 2018), CSPNet (Jiang et al. 2024)) and state-of-the-art self-supervised foundation models (i.e., LaBraM-Base (Jiang, Zhao, and liang Lu 2024), BENDR (Kostas and Rudzicz 2021)). Performance on multi-class tasks (EMO, SLEEP, BCIC-IV 2a) is evaluated using Balanced Accuracy, Cohen's Kappa, and Weighted-F1 score. For the binary classification task (SPMI), Balanced Accuracy, AUC-PR, and AUROC are used. All results are obtained through a 5-fold cross-validation procedure, presenting the mean and standard deviation.

Setup	FACED	ISRUC	BCIC-IV 2a
w/o pretraining	49.70 ± 0.85	66.25 ± 1.20	41.29 ± 1.87
w/o $\mathcal{L}_{\text{recon}}$	56.38 ± 0.90	70.10 ± 1.22	49.37 ± 1.41
w/o $\mathcal{L}_{\text{align}}$	49.17 ± 1.11	64.93 ± 1.14	45.60 ± 1.55
Ours	**59.12 ± 0.85**	**71.86 ± 0.89**	**51.59 ± 1.21**
	DEAP	**HMC**	**SPMI**
w/o pretraining	48.50 ± 1.20	65.12 ± 1.25	62.33 ± 1.35
w/o $\mathcal{L}_{\text{recon}}$	52.15 ± 1.15	68.80 ± 1.10	65.91 ± 1.05
w/o $\mathcal{L}_{\text{align}}$	48.95 ± 1.30	64.75 ± 1.18	62.89 ± 1.28
Ours	**55.82 ± 1.05**	**70.59 ± 1.02**	**68.07 ± 0.95**

Table 3: Ablation study on the effectiveness of the composite pre-training objective. We report Balanced Accuracy (mean ± std). Best results are in bold.

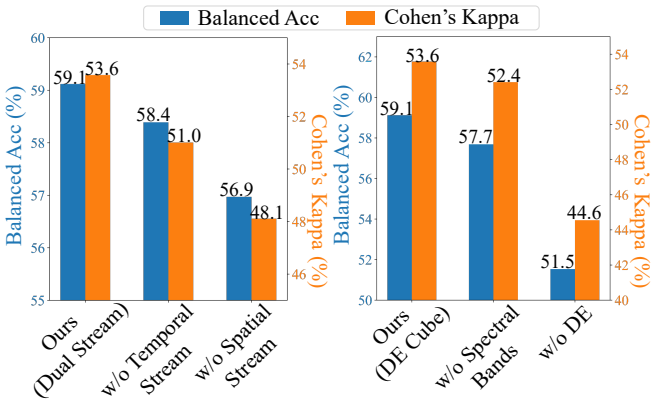

Figure 3: Ablation study on BraSTORM's core components on FACED. **Left:** Comparison of the proposed dual-stream architecture against single-stream baselines. **Right:** Evaluating the effectiveness of the proposed multi-band DE cube.

Performance Comparison As presented in Table 2, BraSTORM (Full Fine-tuning) demonstrates competitive performance, achieving state-of-the-art results on the majority of benchmarks. On the challenging 9-class FACED emotion recognition task, BraSTORM achieves a Balanced Accuracy of 59.12%, which is a notable improvement over the best-performing baseline, EEGPT (57.54%). In sleep staging on the ISRUC dataset, BraSTORM again outperforms all prior models, especially in Kappa (66.92% vs. LaBraM's 63.50%) and Weighted-F1 (73.15% vs. 70.91%), indicating a more reliable classification of sleep stages. For motor imagery on the BCIC-IV 2a dataset, BraSTORM surpasses all foundation model baselines in both Balanced Accuracy and Cohen's Kappa, demonstrating its robust feature extraction capabilities even in complex, low signal-to-noise ratio scenarios.

Ablation Study

Ablation on Pre-training Objectives We first investigate the contributions of each component in our pre-training objective: the reconstruction loss ($\mathcal{L}_{\text{recon}}$) and the alignment loss ($\mathcal{L}_{\text{align}}$). We compare the full BraSTORM model against three variants: training from scratch ("w/o pretraining"), a reconstruction-only model ("w/o $\mathcal{L}_{\text{align}}$"), and a contrastive-only model ("w/o $\mathcal{L}_{\text{recon}}$").

As shown in Table 3, the model trained from scratch performs the worst, confirming the substantial benefit of our self-supervised paradigm. Ablating either loss component individually leads to a significant performance drop across all tasks. Specifically, the sharp decline without $\mathcal{L}_{\text{align}}$ is telling: it demonstrates that even with rich, stream-specific features learned via reconstruction, the model fails to learn the crucial semantic correspondence between the temporal and spatial domains. This alignment loss is therefore indispensable for forcing the two encoders to map features from the same underlying neural event into a unified, coherent embedding space. Conversely, removing $\mathcal{L}_{\text{recon}}$ shows that high-level alignment is insufficient without the fine-grained structural details provided by reconstruction. The synergy between these two objectives is thus critical to the success of BraSTORM.

Ablation on Input Factorization To validate our core hypothesis that explicit input-level factorization is crucial, we compare our dual-stream design against two entangled-input variants. The first, "w/o Spatial Stream", removes the spatial branch entirely, representing temporal-only models. The second, and more critical, variant ("w/o Temporal Stream") replaces our dual-stream architecture with a single ViT encoder that processes a unified, patch-based spatio-temporal input. To ensure a fair comparison, both variants were trained with an adapted composite objective, where the contrastive loss was applied intra-modally between augmented views of their respective single-stream inputs.

The results, shown on the left of Figure 3, provide compelling evidence for our approach. Our dual-stream model significantly outperforms both single-stream variants. The performance drop in the entangled ("w/o Temporal Stream") model is particularly revealing. Despite using the identical pre-training objective, its inability to disentangle spatial and temporal patterns at the input leads to a less effective representation. This strongly suggests that a design-based disentanglement is superior to leaving the task implicitly to a single encoder.

Ablation on Spatial Stream Input We next ablate the components of our spatio-spectral cube to justify its specific design. We compare our full model "Ours (DE Cube)" against two alternatives for the spatial stream: (1) "w/o Spectral Bands", which computes DE features over the entire broadband signal to create a single-channel map ($1 \times H \times W$), and (2) "w/o DE", which replaces Differential Entropy with a basic raw statistics (i.e., the per-channel signal mean) to construct the spatial map.

As shown on the right of Figure 3, our proposed DE-based cube significantly outperforms the alternatives. The reduced performance of the "w/o Spectral Bands" model highlights the importance of frequency-specific information, as different neural oscillations carry distinct functional roles. The

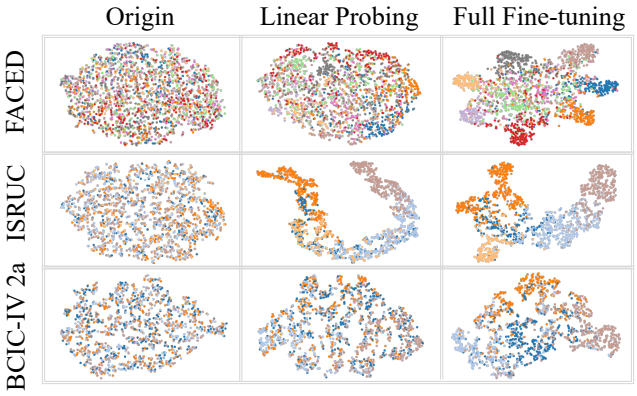

Figure 4: t-SNE visualization of feature distributions. For each dataset, plots from left to right depict the raw data ("Origin"), the fused spatio-temporal representation from the frozen pre-trained encoders ("Linear Probing"), and the same representation after downstream fine-tuning ("Full Fine-tuning").

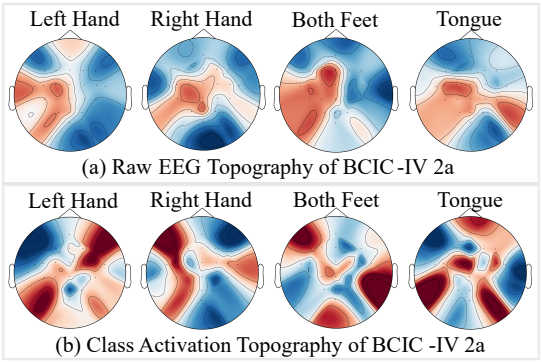

Figure 5: Class activation topography for the BCIC-IV 2a motor imagery task. (a) raw signal power. (b) Class activation maps (CAMs) from BraSTORM's spatial encoder, highlighting brain regions with high contribution to the final class prediction.

further drop when removing DE ("w/o DE") confirms that DE is a more powerful feature than simple statistics for capturing the complexity of underlying neural activity. This validates that the combination of multi-band decomposition and DE features is an effective design for a highly informative spatial representation.

Interpretability Analysis

To provide a more intuitive understanding of BraSTORM's capabilities, two sets of visualization analyses are conducted: one assessing the quality of the learned representation space and another examining the model's neurophysiological interpretability.

Representation Space Analysis with t-SNE We use t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton 2008) to visualize the feature distributions on various downstream datasets, as illustrated in Figure 4. For each state, we visualize the fused spatio-temporal representation, the concatenated output of the temporal (E_T) and spatial (E_S) encoders that is fed to the classifier. Three states are compared: the raw EEG data ("Origin"), the fused representation from the frozen pre-trained encoders ("Linear Probing"), and the final representation after full fine-tuning ("Full Fine-tuning"). The raw data projections show indistinct, overlapping clusters, indicating a lack of inherent class-separable structure. In contrast, the representations from the pre-trained encoder exhibit significantly improved class-wise clustering, even without any task-specific training of the encoder. This demonstrates that our self-supervised pre-training organizes the data into a semantically meaningful manifold. Finally, after full fine-tuning, the clusters become highly compact and well-separated, visually confirming the model's ability to learn discriminative features for downstream tasks.

Interpretability via Class Activation Topography A central claim of this work is that disentangling spatial features allows for more meaningful interpretations. To validate this, Grad-CAM (Selvaraju et al. 2017) is employed to generate class activation topographies for the BCIC-IV 2a motor imagery task, as shown in Figure 5. Specifically, we compute the gradients of the final class prediction with respect to the output patch embeddings of the final block in the spatial encoder (E_S). The resulting maps visualize the contribution of each spatial location to the classification decision. The raw EEG power topography shows diffuse activity with no clear class-specific patterns. In stark contrast, BraSTORM's spatial encoder learns a neurophysiologically plausible localization mechanism. For "Left Hand" and "Right Hand" imagery, the model correctly focuses on the contralateral motor cortex areas known to govern such movements. Furthermore, activations for "Both Feet" and "Tongue" are appropriately concentrated along the medial motor cortex, consistent with the brain's somatotopic organization. This result is a powerful testament to our method's design: by forcing the spatial encoder to learn from purely topographical information, BraSTORM not only improves performance but also discovers interpretable neural signatures directly from the data.

Conclusion

In conclusion, this paper addressed the issue of spatio-temporal entanglement by proposing BraSTORM. More fundamentally, our work establishes a new design principle for self-supervised learning on physiological signals: factorize-then-learn. By demonstrating that an explicit, input-level disentanglement of generative factors is superior to implicit, model-level separation, BraSTORM offers a path toward more robust, generalizable, and interpretable foundation models. This principle has broad implications beyond EEG, potentially informing the design of models for other signals with inherent spatio-temporal structures, such as fNIRS, MEG, or even dynamic graph data. Our future work will explore these exciting frontiers, aiming to build a truly universal foundation model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 72350710798) and the Zhejiang Natural Science Foundation (Grant No. LQN25F020007).

References

Altaheri, H.; Muhammad, G.; and Alsulaiman, M. 2022. Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE transactions on industrial informatics*, 19(2): 2249–2258.

Altaheri, H.; Muhammad, G.; Alsulaiman, M.; Amin, S. U.; Altuwaijri, G. A.; Abdul, W.; Bencherif, M. A.; and Faisal, M. 2023. Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review. *Neural Computing and Applications*, 35(20): 14681–14722.

Alvarez-Estevez, D.; and Rijsman, R. M. 2021. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one*, 16(8): e0256111.

Brunner, C.; Leeb, R.; Müller-Putz, G.; Schlögl, A.; and Pfurtscheller, G. 2008. BCI Competition 2008–Graz data set A. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16(1-6): 1.

Chen, J.; Wang, X.; Huang, C.; Hu, X.; Shen, X.; and Zhang, D. 2023. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10(1): 740.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Chen, Y.; Ren, K.; Song, K.; Wang, Y.; Wang, Y.; Li, D.; and Qiu, L. 2024. EEGFormer: Towards transferable and interpretable large-scale EEG foundation model. *arXiv preprint arXiv:2401.10278*.

Chien, H.-Y. S.; Goh, H.; Sandino, C. M.; and Cheng, J. Y. 2022. Maeeg: Masked auto-encoder for eeg representation learning. *arXiv preprint arXiv:2211.02625*.

Cui, W.; Jeong, W.; Thölke, P.; Medani, T.; Jerbi, K.; Joshi, A. A.; and Leahy, R. M. 2024. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Ding, Y.; Robinson, N.; Zhang, S.; Zeng, Q.; and Guan, C. 2022. TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3): 2238–2250.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Huang, G.; Hu, Z.; Chen, W.; Zhang, S.; Liang, Z.; Li, L.; Zhang, L.; and Zhang, Z. 2022. M3CV: A multi-subject, multi-session, and multi-task database for EEG-based biometrics challenge. *NeuroImage*, 264: 119666.

Hwang, H.-J.; Kim, S.; Choi, S.; and Im, C.-H. 2013. EEG-based brain-computer interfaces: a thorough literature survey. *International Journal of Human-Computer Interaction*, 29(12): 814–826.

Jiang, W.; Zhao, L.; and liang Lu, B. 2024. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *The Twelfth International Conference on Learning Representations*.

Jiang, X.; Meng, L.; Chen, X.; Xu, Y.; and Wu, D. 2024. CSP-Net: Common spatial pattern empowered neural networks for EEG-based motor imagery classification. *Knowledge-Based Systems*, 305: 112668.

Khalighi, S.; Sousa, T.; Santos, J. M.; and Nunes, U. 2016. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124: 180–192.

Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1): 18–31.

Kostas, S.; and Rudzicz, F. 2021. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15: 653659.

Lai, J.; Wei, J.; Yao, L.; and Wang, Y. 2025. A Simple Review of EEG Foundation Models: Datasets, Advancements and Future Perspectives. *arXiv preprint arXiv:2504.20069*.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5): 056013.

Liu, H.; Wei, P.; Wang, H.; Lv, X.; Duan, W.; Li, M.; Zhao, Y.; Wang, Q.; Chen, X.; Shi, G.; et al. 2024a. An EEG motor imagery dataset for brain computer interface in acute stroke patients. *Scientific Data*, 11(1): 131.

Liu, Q.; Zhou, Z.; Wang, J.; and Liang, Z. 2024b. Joint Contrastive Learning with Feature Alignment for Cross-Corpus EEG-based Emotion Recognition. In *Proceedings of the 1st International Workshop on Brain-Computer Interfaces (BCI) for Multimedia Understanding*, 9–17.

Liu, W.; Qiu, J.-L.; Zheng, W.-L.; and Lu, B.-L. 2021. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 715–729.

Mohsenvand, M. N.; Izadi, M. R.; and Maes, P. 2020. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*, 238–253. PMLR.

Phan, H.; and Mikkelsen, K. 2022. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4): 04TR01.

Prabowo, D. W.; Nugroho, H. A.; Setiawan, N. A.; and Debayle, J. 2023. A systematic literature review of emotion recognition using EEG signals. *Cognitive Systems Research*, 82: 101152.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Song, T.; Zheng, W.; Song, P.; and Cui, Z. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541.

Song, Y.; Zheng, Q.; Liu, B.; and Gao, X. 2022. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 710–719.

Subha, D. P.; Joseph, P. K.; Acharya U, R.; and Lim, C. M. 2010. EEG signal analysis: a survey. *Journal of medical systems*, 34: 195–212.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, G.; Liu, W.; He, Y.; Xu, C.; Ma, L.; and Li, H. 2024. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37: 39249–39280.

Wang, H.; Chen, T.; and Song, L. 2024. Cascaded Self-supervised Learning for Subject-independent EEG-based Emotion Recognition. *arXiv preprint arXiv:2403.04041*.

Wang, J.; Zhao, S.; Luo, Z.; Zhou, Y.; Jiang, H.; Li, S.; Li, T.; and Pan, G. 2025. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In *The Thirteenth International Conference on Learning Representations*.

Wang, Y.; Chen, X.; Gao, X.; and Gao, S. 2016. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10): 1746–1752.

Wang, Y.; Zhang, B.; and Tang, Y. 2024. Dmmr: Cross-subject domain generalization for eeg-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 628–636.

Wei, X.; Zhao, K.; Jiao, Y.; Carlisle, N. B.; Xie, H.; and Zhang, Y. 2024. Pre-Training Graph Contrastive Masked Autoencoders are Strong Distillers for EEG. *arXiv preprint arXiv:2411.19230*.

Yang, C.; Westover, M.; and Sun, J. 2023. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36: 78240–78260.

Yang, Y.; Wu, Q.; Fu, Y.; and Chen, X. 2018. Continuous convolutional neural network with 3D input for EEG-based emotion recognition. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part VII 25*, 433–443. Springer.

Zhang, D.; Yuan, Z.; Yang, Y.; Chen, J.; Wang, J.; and Li, Y. 2023. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36: 26304–26321.

Zhang, Z.; Zhong, S.; and Liu, Y. 2024a. Beyond mimicking under-represented emotions: deep data augmentation with emotional subspace constraints for EEG-based emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 10252–10260.

Zhang, Z.; Zhong, S.-h.; and Liu, Y. 2024b. TorchEEGEMO: A deep learning toolbox towards EEG-based emotion recognition. *Expert Systems with Applications*, 249: 123550.

Zheng, W.-L.; Liu, W.; Lu, Y.; Lu, B.-L.; and Cichocki, A. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49(3): 1110–1122.

Zheng, W.-L.; and Lu, B.-L. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3): 162–175.