EEG-DINO: Learning EEG Foundation Models via Hierarchical Self-Distillation

Xujia Wang[1,2][0009−0001−7034−3258], Xuhui Liu[1,2][0000−0001−6064−3401], Xi Liu[1][0000−0003−1149−8951], Qian Si[2][0000−0002−1272−2516], Zhaoliang Xu[2][0009−0005−9680−6272], Yang Li[2][0000−0002−1751−1742], and Xiantong Zhen[1][0000−0001−5213−0462]

[1] Central Research Institue, United Imaging Healthcare, Co., Ltd., Beijing, China
[2] Beihang University, Beijing, China

Abstract. Electroencephalography (EEG) provides a non-invasive window into the brain's electrical activity, playing an essential role in various brain–computer interface (BCI) and healthcare applications. In this paper, we propose EEG-DINO, a novel foundation model for EEG encoding based on a hierarchical self-distillation framework. By multi-view semantic alignment, the model is able to extract multi-level semantic features from EEG data, which captures a wide range of semantic information, increasing the robustness against noise and variances inherent in complex EEG signals. Moreover, acknowledging the unique heterogeneous spatial-temporal dependencies in EEG signals, we design a channel-aware sampling mechanism and a decoupled positional embedding scheme. They independently address spatial and temporal dimensions, enabling the model to capture the intricate structural characteristics of EEG signals. We pre-train EEG-DINO[3] on a large-scale EEG corpus spanning over 9000 hours, which consistently achieves state-of-the-art performance on multiple downstream tasks. These results demonstrate the great effectiveness of our self-distillation framework for EEG encoding.

Keywords: Electroencephalography (EEG) · Self-supervised learning · Foundation models · Self-Distillation · EEG pre-training.

1 Introduction

Electroencephalography (EEG) records electrical activity on the scalp, providing high temporal resolution, making it ideal for real-time brain activity monitoring [15]. EEG is extensively used in the diagnosis and monitoring of neurological disorders, including epilepsy [19], sleep disorders [22], neurodegenerative diseases [3] and neuroprosthetics [1,2]. Effective encoding of EEG signals has emerged as a vital step in effectively performing complex and diverse tasks, as it enables the extraction of meaningful information from raw data, thereby enhancing the accuracy and reliability [17].

[3] The pre-trained weights and code for fine-tuning are anonymously available at: https://huggingface.co/eegdino/EEG-DINO.

Early EEG analysis methods primarily employ deep learning (DL) with various neural networks [9,16,18] to learn signal features. However, they typically rely on supervised learning tailored to specific tasks or datasets, leading to challenges in generalization, especially due to limited data, high noise levels, stable nature over short time intervals, and significant variations in EEG signal formats. Recently, building on advancements of self-supervised learning (SSL) in various fields [6,12], many works [8,20,21] propose EEG foundation models pretrained on large-scale unlabeled EEG data, demonstrating the potential of SSL in learning meaningful spatiotemporal features that generalize across subjects and recording sessions. Nonetheless, these methods usually adopt reconstructive objectives, which are ill-suited to handle the issues of the low signal-to-noise ratio and the stable nature of EEG signals over short time intervals. As a result, they tend to focus on noise reconstruction, limiting the ability of models to capture discriminative neurophysiological patterns and meaningful temporal variations.

In this paper, we propose EEG-DINO, the first distillation-based foundation model for EEG encoding. EEG-DINO adopts a hierarchical self-distillation framework within a shared network architecture, enabling knowledge distillation through self-supervised learning across multiple augmented views. This framework leverages multi-view learning to achieve semantic alignment, enhancing robustness and facilitating the extraction of multi-level semantic features from complex, noisy EEG signals. Moreover, we introduce a channel-aware sampling mechanism and a decoupled positional embedding scheme. They exploit the unique structural characteristics of EEG signals by handling spatial and temporal dimensions separately, enabling effective modeling of the heterogeneous spatial-temporal dependencies inherent in EEG data. We note that our work focuses on EEG foundation model learning based on the self-distillation principles and thereby is fundamentally different from early EEG distillation works [4,5,14], which distill a pre-trained model to a small model for supervised learning on specific tasks. We extensively evaluate EEG-DINO on multiple downstream tasks, where it demonstrates superior performance compared to counterpart methods, setting new state-of-the-arts, highlighting the promise of our distillation-driven self-supervised learning for EEG encoding.

2 Method

This section outlines the hierarchical self-distillation framework with a teacher-student architecture to pre-train EEG-DINO, as shown in Fig. 1. Given the raw EEG signals represented as $X \in \mathbb{R}^{C \times T}$, where C denotes the number of channels and T is the number of timestamps, we first devise a channel-aware sampling mechanism to obtain multi-view inputs, allowing the model to capture multi-level semantic information.

Subsequently, we utilize the time-frequency embedding (TFE) [20] to project the raw signal X to a set of EEG tokens $\mathcal{E}$. Notably, an EEG token is defined as a 1-second temporal segment across all channels. Meanwhile, we introduce the decoupled positional embedding (DPE) scheme to encode spatial and temporal

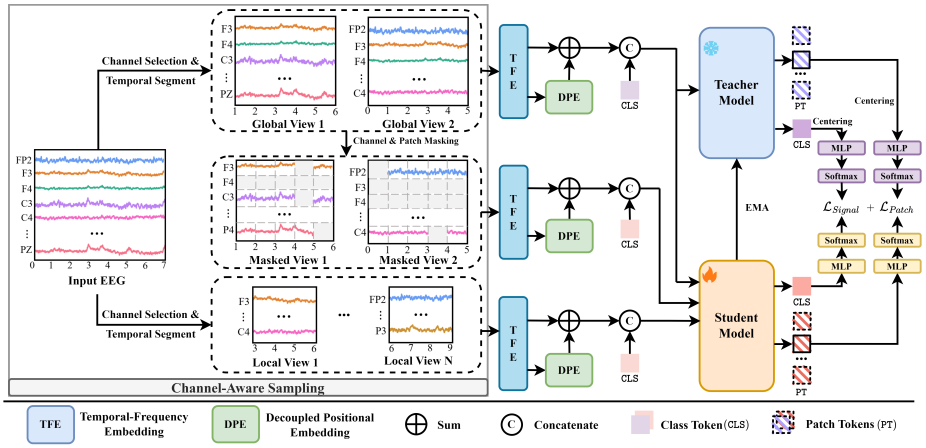

Fig. 1. Overview of the hierarchical self-distillation framework to pre-train EEG-DINO.

positional dependencies separately, which is then integrated to $\mathcal{E}$ to form the input EEG tokens of the transformer for pre-training. In this context, the model learns a set of features that capture temporal and spatial patterns in EEG data. As a result, the pre-trained EEG-DINO can be adapted to conduct specific downstream tasks.

2.1 Tailored Sampling and Positional Embedding for EEG Signals

Channel-Aware Sampling EEG inherently exhibits spatially distributed patterns across sensor channels, where each channel corresponds to localized neural activity in distinct brain regions. To exploit this spatial-temporal characteristic, we design a channel-aware sampling mechanism specifically for EEG to create three view types, forming the hierarchy for self-distillation across views. As shown in Fig. 1, unlike conventional image cropping that extracts contiguous spatial regions, our mechanism uniquely operates on EEG signals by adaptively sampling subsets of channels and temporal segments to construct diverse perspectives. Global views retain a moderate proportion of sparse channels and continuous temporal windows to preserve broad spatial-temporal patterns, local views employ aggressive spatiotemporal reduction to focus on localized dynamics and masked views integrate channel-wise and temporal-patch masking on global views to simulate incomplete observations. This pipeline creates 12 diverse perspectives (2 global, 2 masked, 8 local) per sample to encourages the model to learn robust representations resilient to channel variations, following the perspective configuration in DINO.

Decoupled Positional Embedding After tokenizing multi-views signals using TFE [20], the DPE performs spatial channel encoding and dynamic temporal

encoding separately. The former applies a learnable projection that maps the one-hot channel vector into an embedding $\mathcal{P}_c$. The later dynamically performs channel-wise 1D convolutions along the temporal axis to produce the temporal embedding $\mathcal{P}_t$. This design decouples channel and temporal embeddings, which enables robust generalization across structural characteristics. The final patch embedding combines both components through summation:

$$\text{Embed}(X) = \mathcal{P}_c + \mathcal{P}_t + \mathcal{E} \tag{1}$$

2.2 Pre-Training via Hierarchical Self-Distillation

Teacher-Student Architecture Our framework consists of a teacher model and a student model, instantiating DINO-v2 principles for EEG representation learning. The input sequence for this architecture is constructed by concatenating the class token (`CLS`) with patch embeddings. As shown in Fig. 1, the teacher processes only global views to maintain stable target representations. The student operates on all view types to produce robust feature representations. To ensure stable training, the teacher parameters are updated through exponential moving average (EMA) [7] of the student model parameters. This momentum update mechanism prevents abrupt changes in the target representations while allowing gradual refinement.

Hierarchical Self-Distillation We propose a hierarchical self-distillation framework, where knowledge distillation is performed across views at different levels, which forces the foundation model to learn robust representations.

Signal-level Distillation The distillation at this level applies to `CLS` of teacher from global views and `CLS` of student from global, local and masked views. Following DINO v2 [12], as shown in Fig. 1, `CLS` is passed into an MLP layer followed by a softmax operation to produce probability vector. For the teacher model, we apply centering to the logits (pre-softmax outputs) by subtracting a momentum-updated mean vector to obtain the probability vector. The cross-entropy loss is then computed: $\mathcal{L}_* = -\sum p_t \log p_s$ where $*$ denotes any one of the views, p_t and p_t are the probability vectors from teacher and student models, respectively. We can calculate weighted losses from global, local, and masked views at the signal-level distillation:

$$\mathcal{L}_{Signal} = \mathcal{L}_{Global} + \mathcal{L}_{Local} + \mathcal{L}_{Masked} \tag{2}$$

where $\mathcal{L}_{Global}$ is the cross-entropy loss between teacher and student models for global views; similarly, $\mathcal{L}_{Local}$ and $\mathcal{L}_{Masked}$ are losses for local and masked views, respectively.

Patch-level Distillation This distillation applies to the patch tokens (`PT`) of the teacher model from the global views and the patch tokens `PT` of the student model from masked views [23]. As shown in Fig.1, the respective probability

vectors p_t and p_s from the teacher and student models are obtained in a similar way to compute the loss: $\mathcal{L}_{Patch} = -\sum_i p_t \log p_s$ over all masked patch tokens indexed by i at patch-level distillation.

Total Loss The final total loss for optimization is computed as the average of the two weighted loss components:

$$\mathcal{L} = \mathcal{L}_{Signal} + \mathcal{L}_{Patch} \tag{3}$$

The student model obtained from the pre-training stage will be used as the foundation model to be adapted to downstream tasks by either full-parameter fine-tuning or linear probing.

3 Experiments and Results

3.1 Dataset and Setup

Pre-training Dataset We perform pre-training on the Temple University Hospital EEG (TUEG) dataset [11], one of the largest publicly available EEG datasets. This dataset contains over 30,000 clinical EEG recordings from more than 16,000 patients. Following CBraMod [20], the raw EEG data undergoes several pre-processing steps: we select 19 common channels from the international 10-20 system (e.g., FP1, FP2, F3, F4, etc.), resample all signals to 200 Hz and split them into 30-second epochs. Finally, a total of 1,109,545 EEG samples, over 9000 hours in duration, are retained for pre-training.

Downstream Datasets To demonstrate the generalization of our model, we systematically evaluate our model on several different types of downstream datasets:

TUEV [11] is a curated subset of the TUEG database annotated for six typical events. Following protocols from BIOT [21], EEG signals recorded via 19 standardized bipolar electrode pairs (10-20 system) were downsampled to 200 Hz, and segmented into 112,491 5-second epochs. Training subjects were split into an 80% training set and a 20% validation set.

TUAB [11] represents a curated selection from the TUEG database and have been annotated as normal or abnormal. We have used the same preprocessing method as TUEV. Finally, all EEG signals are resampled to 200 Hz and divided into 409,455 10-second 19-channel samples.

SEED-V [10] is a multimodal emotion recognition dataset featuring five emotions. It uses EEG recordings from 16 participants (open-source version) across three sessions with 15 trials each. EEG signals, originally recorded at 1000 Hz using 62 electrodes and downsampled to 200 Hz, were segmented into 117,744 one-second segments. Each session's trials are evenly divided (5:5:5) into three subsets for balanced experimental validation, following CBraMod's setup [20].

Environments and Settings The experiments are implemented by Python 3.11.11, Torch 2.5.1+cu124, on eight H800 GPUs. All the models are optimized on training set, selected from the validation set and evaluated on the test set. We obtain five sets of results with different random seeds and report the mean and standard deviation values. We devise three different configurations: EEG-DINO-(S)mall, EEG-DINO-(M)edium , and EEG-DINO-(L)arge as shown in Table 1.

Table 1. Hyperparameters for EEG-DINO pre-training

Model	Layers	Hidden Size	MLP Size	Params
EEG-DINO-S	12	200	512	4.6M
EEG-DINO-M	16	512	1024	33M
EEG-DINO-L	24	1024	2048	201M

Baselines & Metrics The baselines of foundation models are from BIOT [21], LaBraM-(B)ase [8], CBraMod [20] and the non-foundation models baselines are from CNN-(T)ransformer [13] and ST-(T)ransformer [16]. To make head-to-head comparisons, we fine-tune BIOT, LaBraM-B and CBraMod based on their public code, pre-trained weights and public parameter settings under the same dataset seeds as ours. We employ Balanced Accuracy (BA), AUC-PR and AUROC as evaluation metrics for binary classification. And for multi-class classification, we employ BA, Cohen's Kappa and Weighted F1 Score. For model optimization and selection, we designate AUROC as the monitoring metric for binary classification tasks and Cohen's Kappa for multi-class classification tasks.

3.2 Results

We performed experiments comparing linear probing and full-parameter fine-tuning, where full-parameter fine-tuning updates all model weights including the backbone and the added classification head, and linear probing updates classification head only, keeping backbone parameters fixed.

Linear Probing Tables 2 and 3 compare EEG-DINO-S/M/L with state-of-the-art self-supervised EEG models across three benchmark datasets using a linear probing protocol. EEG-DINO-S significantly outperforms counterpart methods on multiple datasets (TUEV, SEED-V, TUAB) in terms of all metrics, while using fewer or similar parameters. Our EEG-DINO performs exceptionally well even without full-parameter fine-tuning, which demonstrate that EEG-DINO as a foundational model has a strong encoding capability. Furthermore, increasing model capacity (from Small to Medium and Large) consistently enhances EEG representation learning, suggesting that scaling data and parameters leads to more generalized pattern extraction for diverse downstream EEG tasks.

Table 2. The results of linear probing on TUEV and SEED-V

Dataset	Methods	Params	BA(%)	Cohen's Kappa	Weighted F1
TUEV	BIOT [21]	3.2M	33.27 ± 2.56	0.3835 ± 0.0554	0.6792 ± 0.0288
	LaBraM-B [8]	5.8M	34.61 ± 2.25	0.3968 ± 0.0329	0.6974 ± 0.0161
	CBraMod [20]	4.0M	32.46 ± 2.72	0.3884 ± 0.1824	0.6889 ± 0.0625
	EEG-DINO-S	4.6M	54.82 ± 1.06	0.5673 ± 0.0023	0.7861 ± 0.0024
	EEG-DINO-M	33M	58.80 ± 0.68	0.6180 ± 0.0145	0.8111 ± 0.0066
	EEG-DINO-L	201M	$\mathbf{60.54} \pm 0.53$	$\mathbf{0.6419} \pm 0.0122$	$\mathbf{0.8214} \pm 0.0045$
SEED-V	BIOT [21]	3.2M	24.61 ± 2.87	0.0798 ± 0.0361	0.2489 ± 0.0257
	LaBraM-B [8]	5.8M	25.21 ± 2.67	0.0854 ± 0.0342	0.2543 ± 0.0265
	CBraMod [20]	4.0M	25.36 ± 2.57	0.0842 ± 0.0384	0.2568 ± 0.0275
	EEG-DINO-S	4.6M	29.81 ± 0.35	0.1273 ± 0.0052	0.3035 ± 0.0063
	EEG-DINO-M	33M	33.65 ± 0.56	0.1707 ± 0.0047	0.3426 ± 0.0052
	EEG-DINO-L	201M	$\mathbf{35.79} \pm 0.33$	$\mathbf{0.1984} \pm 0.0029$	$\mathbf{0.3652} \pm 0.0041$

Table 3. The reults of linear probing on TUAB

Methods	Params	BA(%)	AUC-PR	AUROC
BIOT [21]	3.2M	73.08 ± 0.29	0.7849 ± 0.0036	0.8013 ± 0.0058
LaBraM-B [8]	5.8M	74.57 ± 0.14	0.8081 ± 0.0009	0.8115 ± 0.0014
CBraMod [20]	4.0M	67.85 ± 1.33	0.7721 ± 0.0259	0.7826 ± 0.0452
EEG-DINO-S	4.6M	78.41 ± 0.08	0.8666 ± 0.0006	0.8706 ± 0.0004
EEG-DINO-M	33M	79.15 ± 0.11	0.8680 ± 0.0008	0.8763 ± 0.0010
EEG-DINO-L	201M	$\mathbf{79.63} \pm 0.07$	$\mathbf{0.8701} \pm 0.0014$	$\mathbf{0.8814} \pm 0.0007$

Fine-Tuning As shown in Tables 4 and 5, our experiments demonstrate that EEG-DINO-S outperforms the baselines in all metrics on the TUEV and the SEED-V datasets, showing enhanced inter-subject agreement and better generalization. In binary detection on the TUAB dataset, our model performs similarly to the baselines while utilizing fewer parameters, highlighting its efficiency in maintaining high performance with reduced complexity. Scaling to EEG-DINO-M/L further boosts performance. Additionally, we use a randomly initialized model (EEG-DINO-S*) for fine-tuning as a contrast to the pre-trained version. It underperforms compared to the pre-trained versions, which further verifies the effectiveness of our pre-training approach. While it still exceeds supervised CNN-Transformer baselines, underscoring the architecture's effectiveness.

Ablation Study To further evaluate the effectiveness of our proposed techniques, we conduct three ablation studies on pre-training: w/o decoupled positional embedding (DPE), random masking strategy and full cropping strategy. The results are presented in Fig. 2, which demonstrates that all three techniques are effective across the three datasets used. This indicates that these techniques are well-suited for EEG signals and can exploit the unique structural characteristics, enhancing our model's performance and robustness.

Table 4. The results of fine-tuning on TUEV and SEED-V

Dataset	Methods	Params	BA(%)	Cohen's Kappa	Weighted F1
TUEV	CNN-T [13]	3.2M	40.87 ± 1.61	0.3815 ± 0.0134	0.6854 ± 0.0293
	ST-T [16]	3.5M	39.84 ± 2.28	0.3765 ± 0.0306	0.6823 ± 0.0190
	BIOT [21]	3.2M	52.81 ± 2.25	0.5273 ± 0.0249	0.7492 ± 0.0082
	LaBraM-B [8]	5.8M	64.09 ± 0.65	0.6637 ± 0.0093	0.8312 ± 0.0052
	CBraMod [20]	4.0M	59.42 ± 1.32	0.5818 ± 0.0149	0.7817 ± 0.0201
	EEG-DINO-S*	4.6M	50.20 ± 1.74	0.4620 ± 0.3960	0.7307 ± 0.2072
	EEG-DINO-S	4.6M	65.16 ± 0.64	0.6654 ± 0.0082	0.8356 ± 0.0046
	EEG-DINO-M	33M	66.11 ± 0.95	0.6739 ± 0.0123	0.8357 ± 0.0054
	EEG-DINO-L	201M	$\mathbf{66.79} \pm 0.57$	$\mathbf{0.6809} \pm 0.0094$	$\mathbf{0.8398} \pm 0.0049$
SEED-V	CNN-T [13]	3.2M	36.78 ± 0.78	0.2072 ± 0.0183	0.3642 ± 0.0088
	ST-T [16]	3.5M	30.52 ± 0.72	0.1083 ± 0.0121	0.2833 ± 0.0105
	BIOT [21]	3.2M	38.37 ± 1.87	0.2261 ± 0.0262	0.3856 ± 0.0203
	LaBraM-B [8]	5.8M	39.76 ± 1.38	0.2386 ± 0.0209	0.3974 ± 0.0111
	CBraMod [20]	4.0M	38.99 ± 0.25	0.2414 ± 0.0013	0.3977 ± 0.0058
	EEG-DINO-S*	4.6M	38.83 ± 0.33	0.2399 ± 0.0046	0.3963 ± 0.0036
	EEG-DINO-S	4.6M	40.63 ± 0.45	0.2564 ± 0.0067	0.4092 ± 0.0060
	EEG-DINO-M	33M	41.38 ± 0.32	0.2727 ± 0.0055	0.4234 ± 0.0052
	EEG-DINO-L	201M	$\mathbf{41.77} \pm 0.40$	$\mathbf{0.2801} \pm 0.0051$	$\mathbf{0.4315} \pm 0.0042$

* Random initialization before fine-tuning.

Table 5. The reults of fine-tuning on TUAB

Methods	Params	BA(%)	AUC-PR	AUROC
CNN-T [13]	3.2M	77.77 ± 0.22	0.8433 ± 0.0039	0.8461 ± 0.0013
ST-T [16]	3.5M	79.66 ± 0.23	0.8521 ± 0.0026	0.8707 ± 0.0019
BIOT [21]	3.2M	79.59 ± 0.57	0.8792 ± 0.0023	0.8815 ± 0.0043
LaBraM-B [8]	5.8M	81.40 ± 0.19	0.8965 ± 0.0016	0.9022 ± 0.0009
CBraMod [20]	4.0M	80.91 ± 0.32	0.8906 ± 0.0018	0.8831 ± 0.0030
EEG-DINO-S*	4.6M	79.17 ± 0.16	0.8662 ± 0.0017	0.8708 ± 0.0021
EEG-DINO-S	4.6M	81.37 ± 0.36	0.8906 ± 0.0021	0.8981 ± 0.0016
EEG-DINO-M	33M	81.55 ± 0.32	0.8963 ± 0.0011	0.9018 ± 0.0017
EEG-DINO-L	201M	$\mathbf{82.07} \pm 0.24$	$\mathbf{0.9012} \pm 0.0015$	$\mathbf{0.9100} \pm 0.0009$

* Random initialization before fine-tuning.

4 Conclusion

This paper introduces EEG-DINO, a novel distillation-driven self-supervised learning framework for EEG signal encoding. By integrating hierarchical self-distillation with multi-view semantic alignment, channel-aware sampling, and decoupled positional embedding, our approach effectively captures discriminative spatiotemporal EEG signal patterns. Extensive evaluations demonstrate superior performance of EEG-DINO over counterparts. Our work underscores the

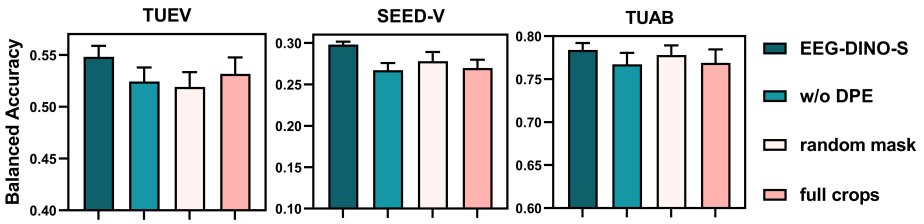

Fig. 2. The results of ablation study on TUEV, SEED-V and TUAB (linear probing)

potential of self-supervised knowledge distillation for robust EEG representation learning, paving the way for adaptive, scalable neurotechnology solutions.

Acknowledgments. This study was partially funded by the National Natural Science Foundation of China (Grant No. 62176068, 623B2011, 62325301 and U24B20186).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Saegh, A., Dawwd, S.A., Abdul-Jabbar, J.M.: Deep learning for motor imagery eeg-based classification: A review. Biomedical Signal Processing and Control **63**, 102172 (2021)
2. Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S.U., Altuwaijri, G.A., Abdul, W., Bencherif, M.A., Faisal, M.: Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. Neural Computing and Applications **35**(20), 14681–14722 (2023)
3. Babiloni, C., Arakaki, X., Azami, H., Bennys, K., Blinowska, K., Bonanni, L., Bujan, A., Carrillo, M.C., Cichocki, A., de Frutos-Lucas, J., et al.: Measures of resting state eeg rhythms for clinical trials in alzheimer's disease: recommendations of an expert panel. Alzheimer's & Dementia **17**(9), 1528–1553 (2021)
4. Fan, C., Zhang, H., Huang, W., Xue, J., Tao, J., Yi, J., Lv, Z., Wu, X.: Dgsd: Dynamical graph self-distillation for eeg-based auditory spatial attention detection. Neural Networks **179**, 106580 (2024)
5. Ferrante, M., Boccato, T., Bargione, S., Toschi, N.: Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models. Computers in Biology and Medicine **178**, 108701 (2024)
6. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
8. Jiang, W.B., Zhao, L.M., Lu, B.L.: Large brain model for learning generic representations with tremendous eeg data in bci. arXiv preprint arXiv:2405.18765 (2024)

9. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)
10. Liu, W., Qiu, J.L., Zheng, W.L., Lu, B.L.: 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 (2021)
11. Obeid, I., Picone, J.: The temple university hospital eeg data corpus. Frontiers in neuroscience **10**, 196 (2016)
12. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
13. Peh, W.Y., Yao, Y., Dauwels, J.: Transformer convolutional neural networks for automated artifact detection in scalp eeg. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 3599–3602. IEEE (2022)
14. Peng, R., Du, Z., Zhao, C., Luo, J., Liu, W., Chen, X., Wu, D.: Multi-branch mutual-distillation transformer for eeg-based seizure subtype classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering **32**, 831–839 (2024)
15. Ranjan, R., Sahana, B.C., Bhandari, A.K.: Deep learning models for diagnosis of schizophrenia using eeg signals: emerging trends, challenges, and prospects. Archives of Computational Methods in Engineering **31**(4), 2345–2384 (2024)
16. Song, Y., Jia, X., Yang, L., Xie, L.: Transformer-based spatial-temporal feature learning for eeg decoding. arXiv preprint arXiv:2106.11170 (2021)
17. Sun, J., Shen, A., Sun, Y., Chen, X., Li, Y., Gao, X., Lu, B.: Adaptive spatiotemporal encoding network for cognitive assessment using resting state eeg. npj Digital Medicine **7**(1), 375 (2024)
18. Supakar, R., Satvaya, P., Chakrabarti, P.: A deep learning based model using rnn-lstm for the detection of schizophrenia from eeg data. Computers in Biology and Medicine **151**, 106225 (2022)
19. Tasci, I., Tasci, B., Barua, P.D., Dogan, S., Tuncer, T., Palmer, E.E., Fujita, H., Acharya, U.R.: Epilepsy detection in 121 patient populations using hypercube pattern from eeg signals. Information Fusion **96**, 252–268 (2023)
20. Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., Pan, G.: Cbramod: A criss-cross brain foundation model for eeg decoding. arXiv preprint arXiv:2412.07236 (2024)
21. Yang, C., Westover, M., Sun, J.: Biot: Biosignal transformer for cross-data learning in the wild. Advances in Neural Information Processing Systems **36**, 78240–78260 (2023)
22. Zhao, W., Van Someren, E.J., Li, C., Chen, X., Gui, W., Tian, Y., Liu, Y., Lei, X.: Eeg spectral analysis in insomnia disorder: A systematic review and meta-analysis. Sleep medicine reviews **59**, 101457 (2021)
23. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)