One Model for All: Universal Pre-training for EEG based Emotion Recognition across Heterogeneous Datasets and Paradigms

Xiang Li*, You Li*, and Yazhou Zhang

***Abstract*—Electroencephalogram (EEG)-based emotion recognition holds significant promise but is hampered by the profound heterogeneity across datasets, particularly in channel configurations and subject variability, which hinders the development of generalizable models. Existing deep learning approaches often require dataset-specific designs and struggle to transfer knowledge effectively. To address this, we propose 'One Model for All', a novel universal pre-training framework designed for EEG analysis across disparate datasets and paradigms. Our core paradigm decouples representation learning into two stages: (1) Univariate pre-training using self-supervised contrastive learning on individual channels, enabled by a proposed Unified Channel Schema (UCS) that leverages the union of channels to maximize data utilization across datasets with varying layouts (e.g., SEED-62ch, DEAP-32ch); (2) Multivariate fine-tuning using a novel spatio-temporal architecture, combining an 'ART' (Adaptive Resampling Transformer) encoder with a 'GAT'(Graph Attention Network), to capture complex spatio-temporal dependencies for downstream tasks. Comprehensive experiments demonstrate the remarkable effectiveness of our approach. Universal pre-training is proven to be an essential stabilizer and performance booster, preventing training collapse on SEED (vs. failure from scratch) and yielding substantial gains on DEAP (+7.65%) and DREAMER (+3.55%). Our framework achieves new state-of-the-art (SOTA) performance across all within-subject benchmarks: 'SEED (99.27%)', 'DEAP (93.69%)', and 'DREAMER (93.93%)'. We further demonstrate SOTA performance on cross-dataset transfer tasks, with our model achieving '94.08%' (via channel intersection) and '93.05%' (via UCS) on the unseen DREAMER dataset, with the former surpassing the within-domain pre-training benchmark. In-depth ablation studies validate our architectural choices, revealing that the GAT module is critical for SOTA performance, yielding a '+22.19%' gain over GCN on the high-noise DEAP dataset, while the removal of the graph network causes a catastrophic '16.44%' performance drop. This work paves the way for building more universal, scalable, and effective pre-trained models for diverse EEG analysis tasks.**

***Index Terms*—Electroencephalogram (EEG), Emotion Recognition, Self-supervised Learning, Pre-training, Transfer Learning, Heterogeneous Datasets.**

Corresponding author: Xiang Li, Yazhou Zhang.

*Xiang Li and You Li make equal contribution to this work and share the co-first authorship.

Xiang Li and You Li are with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center inJinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China and Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China. (e-mail: xiangli@sdas.org).

Yazhou Zhang is with the College of Intelligence and Computing, Tianjin University, Tianjin, China. (e-mail: yzhou_zhang@tju.edu.cn).

I. Introduction

THE advancement of deep learning for Electroencephalogram (EEG)-based emotion recognition is critically hindered by 'profound data heterogeneity' [1]. Datasets vary significantly in experimental paradigms, stimulus types, and technical specifications, particularly the number and layout of recording channels (e.g., 62 channels in SEED [2], 32 in DEAP [3], 14 in DREAMER [4]). Compounding this is the 'high inter-subject variability' inherent in EEG signals, where patterns for the same emotion differ drastically between individuals [5]. These inconsistencies make it extremely difficult for conventional end-to-end models to generalize or transfer knowledge effectively, fragmenting the field into specialized, single-dataset models with limited broader applicability [1]. This lack of generalizability prevents the development of robust, large-scale foundational models for diverse EEG analysis tasks.

While self-supervised pre-training offers a promising path to leverage unlabeled data, mirroring successes in NLP [6] and CV [7], its application to heterogeneous EEG data remains challenging. Existing time series pre-training methods [8], whether based on contrastive learning [9] or mask-reconstruction [7], often do not explicitly address the critical channel mismatch problem across datasets or struggle to rival the performance of task-specific supervised models after fine-tuning.

To overcome these limitations, we propose 'One Model for All,' a novel framework centered around universal pre-training specifically designed for EEG based emotion recognition across heterogeneous datasets. Our core insight is to 'decouple representation learning', first learning universal temporal features through 'univariate pre-training' on individual channels using contrastive self-supervised learning [10]. This stage is crucially enabled by our proposed 'Unified Channel Schema (UCS)', a novel method to handle varying channel layouts by creating a global channel vocabulary from the union of channels, maximizing data utilization. Subsequently, a 'multivariate fine-tuning' stage employs the pre-trained encoder within a novel spatio-temporal architecture. This architecture combines our proposed 'ART' (Adaptive Resampling Transformer), a superior temporal encoder that adaptively handles varying temporal resolutions, with a 'GAT' (Graph Attention Network) to explicitly model inter-channel dependencies. This architectural choice is critical: our experiments demonstrate that while modeling multi-variable channel relationships is essential (a

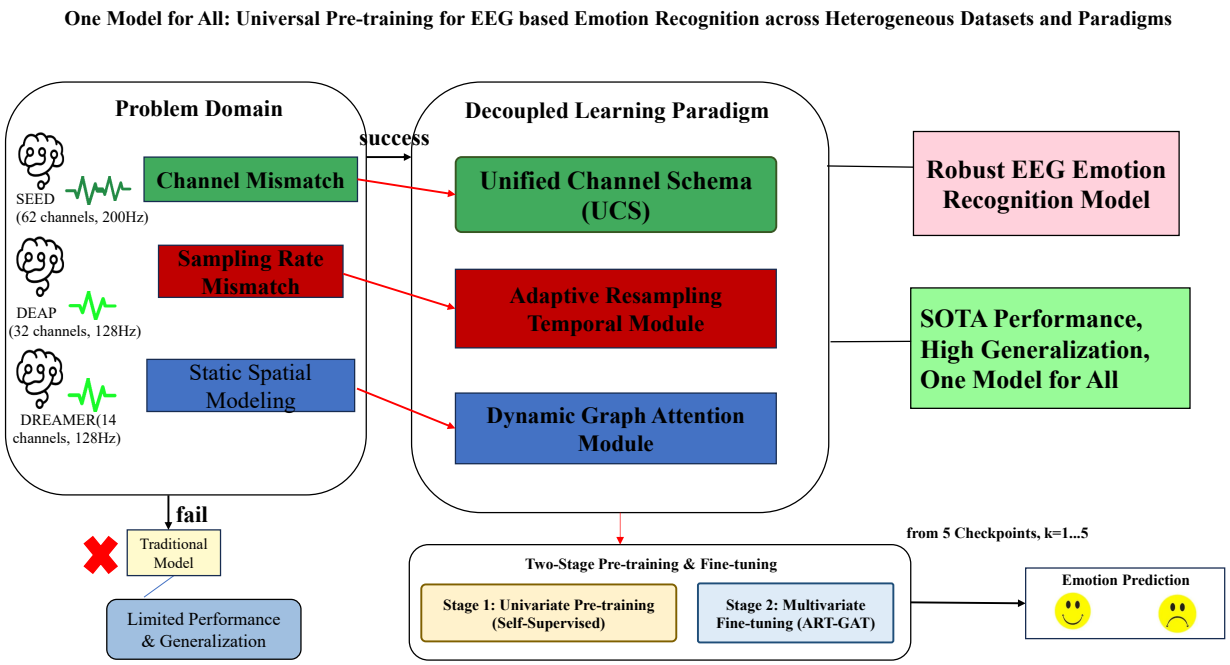


Fig. 1: A conceptual overview of our proposed framework. **(Left)** The core challenge: EEG datasets are highly heterogeneous in channels, sampling rates, and subjects. **(Right)** Our solution: A universal pre-training framework designed to learn a unified representation from these disparate sources.

'16.44%' performance drop without it on DEAP), the attention mechanism of GAT is uniquely capable of handling the high-noise signals where simpler GCNs fail (a '+22.19%' gain on DEAP).

This decoupled approach, leveraging a powerful new architecture, simplifies the learning process at each stage and promotes the development of more generalizable temporal representations before tackling complex spatial interactions.

Our main contributions are threefold:

1) We propose a novel 'decoupled pre-training and fine-tuning paradigm' designed for heterogeneous EEG data, featuring the 'Unified Channel Schema (UCS)' to effectively handle varying channel configurations across datasets.
2) We are the first, to our knowledge, to successfully demonstrate effective 'universal pre-training for emotion recognition across multiple standard, yet highly heterogeneous, EEG datasets' (SEED [2], DEAP [3], DREAMER [4]), enabling significant knowledge transfer even to unseen datasets.
3) We conduct comprehensive experiments validating our new framework's effectiveness. We achieve new state-of-the-art (SOTA) performance across all within-subject benchmarks: SEED (99.27%), DEAP (93.69%), and DREAMER (93.93%). We also demonstrate SOTA performance on cross-dataset transfer, achieving '94.08%' on the unseen DREAMER dataset. Our in-depth ablation studies validate this new architecture, proving that: (a) our pre-training paradigm is an essential 'stabilizer' that prevents training collapse (vs. failure on SEED) and provides significant gains (e.g., '+7.65%' on DEAP vs. scratch); and (b) the 'GAT' module is the critical component for handling high-noise data, yielding a '+22.19%' gain over a standard GCN on DEAP.

The rest of the paper is organized as follows: Section II reviews related literature. Section III details our proposed framework. Section IV presents the experimental setup, results, and analyses. Finally, Section V concludes the paper.

II. Related Works

A. Supervised EEG Emotion Recognition

Supervised learning has been the conventional approach for EEG-based emotion recognition, broadly categorized into feature-based methods and end-to-end deep learning models. Traditional methods first rely on manually extracting features from various domains. These include time-domain (e.g., Hjorth parameters [11]), frequency-domain (e.g., Power Spectral Density, PSD [2]), and time-frequency features (e.g., wavelet transforms [12]). Differential Entropy (DE) has been shown to be a particularly effective and stable feature for emotion recognition [13]. Some methods also engineer features based on functional connectivity (FC), such as the Phase-Locking Value (PLV) [14]. These handcrafted features are then fed into classical machine learning classifiers like Support Vector Machines (SVM) [15], k-Nearest Neighbors (k-NN) [16], or Random Forests [17]. In recent years, end-to-end deep learning models have become prevalent due to their ability to automatically learn discriminative representations [1]. Architectures such as Convolutional Neural Networks (CNNs) [18], Recurrent Neural Networks (RNNs) [19], and their variants like

Long Short-Term Memory (LSTM) [20] and Gated Recurrent Units (GRU) [21] have been widely adopted. More advanced architectures have been proposed to specifically model the complex nature of EEG data. These include Densely Connected Convolutional Networks (DenseNet) [22] for feature reuse, Graph Neural Networks (GNNs) [23] to explicitly model spatial relationships between electrodes, and Transformer models [24] to capture long-range temporal dependencies. Hybrid models combining these architectures, such as CNN-Transformer [25], GNN-Transformer [26], and CNN-GRU [27], have also shown promise. Despite high accuracies in subject-dependent or single-dataset scenarios, supervised models suffer from poor generalization [1]. This limitation stems from profound data heterogeneity, including high inter-subject variability and significant 'domain shifts' across different datasets (cross-corpus) [28]. These shifts are caused by differences in subjects, experimental protocols, stimuli, recording equipment, and channel layouts [29]. Traditional domain adaptation (DA) techniques, such as DANN [30], attempt to align feature distributions but often fail when this domain gap is too large [31]. Furthermore, a lack of standardized evaluation protocols complicates fair comparison across studies [29].

B. Self-Supervised Pre-training for Time Series

To address the label scarcity and generalization issues inherent in supervised learning, Self-Supervised Learning (SSL) has emerged as a powerful paradigm [32]. SSL enables models to learn robust representations from large amounts of unlabeled data by solving pretext tasks [9]. These tasks are diverse, with common frameworks categorized as predictive, generative, and contrastive [9]. Generative models, such as Autoencoders (AEs) [33], Generative Adversarial Networks (GANs) [34], and Diffusion Models (DMs) [35], learn representations by reconstructing the input data. A dominant generative approach is the Masked Autoencoder (MAE) [7], inspired by BERT in NLP [6], which masks portions of the input and trains the model to reconstruct them. This has been adapted for time series [36] and graph data, e.g., GraphMAE [37]. However, applying this directly to EEG is challenging. Due to the low signal-to-noise ratio (SNR) of EEG, reconstruction-based objectives may be suboptimal, as they might force the model to expend capacity on modeling spurious noise and subject-specific artifacts rather than robust neural features [32]. In contrast, Contrastive Learning (CL) learns representations by maximizing the similarity between different augmented 'views' of the same sample (a positive pair) while minimizing their similarity to other samples (negative pairs) [9]. Frameworks like SimCLR [38], MoCo [39], and the InfoNCE loss [40] are foundational. This paradigm is well-suited for learning representations that are invariant to specified augmentations, making it a popular choice for EEG-SSL [10]. Hybrid approaches that combine the strengths of both generative and contrastive learning are also beginning to emerge [41].

C. Self-Supervised Learning in EEG

Applying the general paradigms of Self-Supervised Learning (SSL) to the unique challenges of the EEG domain has given rise to specialized solution paths.In the EEG domain, SSL approaches have primarily evolved along two lines: time-series-based methods and graph-based methods. Some research has focused on combining SSL with supervised signals. For instance, Li et al. [42] proposed SI-CLEER, a **joint learning** framework that concurrently optimizes both a self-supervised contrastive learning loss and a supervised classification loss. This method leverages label information to enhance the contrastive process and demonstrates strong performance on the **single, homogeneous** SEED dataset. However, such joint-training approaches are designed for a single-dataset setting and do not address the challenge of pre-training across **heterogeneous** datasets with different channel layouts. Another line of research converts EEG signals into graph structures, applying Graph Neural Networks (GNNs) and graph-based SSL techniques. For example, Wei et al. [43] proposed EEG-DisGCMAE, a complex framework unifying Graph Contrastive Learning (GCL) and Graph Masked Autoencoders (GMAE), combined with Graph Knowledge Distillation (GKD). This approach transfers knowledge from a **high-density (HD)** 'teacher' model to a **low-density (LD)** 'student' model. While this addresses the HD-LD adaptation problem, its method for unifying different datasets involves **downsampling** the high-density data to match the low-density configuration. The authors note that this approach '**leads to a loss of information**' and that unifying different EEG systems remains a '**key challenge**'.

Our 'One Model for All' framework is designed specifically to solve this heterogeneity challenge **without information loss**. In contrast to the *joint-training* model of SI-CLEER or the complex *graph-distillation* model of EEG-DisGCMAE, we propose a **decoupled two-stage paradigm** (detailed in **Section III-A**). We first learn universal features via **univariate pre-training** on individual channels (**Section III-B**). Crucially, to solve the channel mismatch, we introduce the **Unified Channel Schema (UCS)** (**Section III-B3**). Unlike methods that unify datasets by taking a channel intersection (a subset), which leads to information loss, the UCS utilizes the union of all available channels from heterogeneous datasets (e.g., SEED-62ch and DEAP-32ch) to create a global vocabulary. This allows our model to maximize data utilization and provides a truly scalable solution (validated in **Section IV-B1**) for integrating the diverse EEG sources that prior works found challenging.

III. Proposed Framework

A. The Core Paradigm: Decoupled Univariate Pre-training and Multivariate Fine-tuning

Developing a general-purpose, pre-trained model for electroencephalogram (EEG) analysis is a significant goal, yet it is hindered by profound heterogeneity across datasets. For instance, the SEED dataset includes 62 channels [2], while the DEAP dataset includes 32 [3]. This variability in input channels is compounded by differences in sampling rates and the significant inter-subject variability inherent to EEG signals. While channel heterogeneity is addressed by our Unified Channel Schema (UCS,Section III-B3), handling sampling

rate differences typically requires destructive pre-processing (i.e., uniform downsampling). These inconsistencies pose a fundamental challenge to conventional end-to-end models, which typically require a fixed input dimensionality.

To overcome these limitations, we propose a learning paradigm founded on the principle of decoupling. Rather than tasking a single model with the simultaneous optimization of intra-channel temporal dynamics and inter-channel spatial dependencies, our framework strategically decomposes representation learning into two sequential, specialized stages: **univariate pre-training** and **multivariate fine-tuning**—a design specifically tailored to address the practical complexities of cross-domain learning with different settings and paradigms. This paradigm is instantiated in our work by: (1) pre-training our novel ART (Adaptive Resampling Transformer) encoder on individual channels, which is designed to handle temporal (sampling rate-related) and spatial (channel dimention-related) heterogeneity at the architectural level, and (2) fine-tuning a GAT (Graph Attention Network) classifier to learn the complex, personalized spatial relationships. This decoupled methodology offers the following advantages that directly resolving key pain points of source-target domain mismatches.

- It inherently accommodates the **heterogeneity of source-domain datasets**: within the source domain, individual datasets often exhibit inconsistent attributes (e.g., varying channel dimensions, divergent sensor configurations). Since univariate pre-training focuses exclusively on modeling intra-channel temporal dynamics (without relying on inter-channel spatial correlations), it can independently adapt to each dataset's unique channel-level properties—avoiding the conflicts that arise when forcing a single model to align disparate source-domain structures in one step.
- It enables flexibility for the **uncertainty of target-domain output dimensions**: the target domains are often variable and lack pre-definable output dimensions. By deferring the learning of inter-channel spatial dependencies to the multivariate fine-tuning stage, the framework can dynamically adjust the output layer's dimension to match target-domain needs. This eliminates the rigidity of fixed output structures, ensuring the model remains applicable even when target-domain specifications are not known in advance.

B. Stage One: Foundational Encoder Pre-training via Contrastive Learning

The initial stage of our framework is dedicated to learning universal, subject-agnostic EEG representations. This is accomplished by pre-training a Transformer-based encoder on a comprehensive dataset from all subjects using a self-supervised contrastive learning objective.

1) Backbone Architecture and Objective Modification: Our model is built upon a standard Transformer architecture [24], chosen for its proven ability to capture long-range dependencies in sequential data. The primary goal of EEG representation learning, however, is not high-fidelity signal reconstruction but the extraction of stable, task-relevant neural signatures from inherently non-stationary signals with a low signal-to-noise ratio (SNR). Reconstruction-based objectives, as employed by Masked Autoencoders (MAEs) [7], are suboptimal for this task, as they often compel the model to allocate capacity to modeling spurious noise and subject-specific idiosyncrasies rather than core neural features. We therefore pivot from a reconstruction paradigm to a representation-level contrastive learning framework [9]. The rationale is to prioritize representational consistency over signal fidelity.

The contrastive learning based approach drives the model to map different augmented instances of the same cognitive event to neighboring points in the embedding space. This forces the model's representations to become invariant to nuisances like noise and minor temporal shifts, while preserving essential semantic content—with the ultimate goal of directly extracting robust and generalizable neural markers of cognitive states.

2) Self-Supervised Contrastive Task: Our self-supervised task is formulated as an instance-wise contrastive learning problem, aiming to learn instance-discriminative representations. The core objective is to maximize the similarity between two augmented views of the same input sample (a positive pair) while minimizing their similarity to all other samples in a given batch (negative pairs). For an input EEG segment x_i, two augmented views, x_i^a and x_i^b, are generated using a Random Resized Cropping strategy. Specifically, a subsequence of random length is cropped from the input signal and subsequently resized to the original length via linear interpolation. This augmentation is highly effective for time-series data as it forces the model to learn representations that are invariant to variations in temporal scaling and shifts, which are common in physiological signals [8]. The encoder network, denoted as f_θ, maps these augmented views into a latent representation space and is trained to minimize the NT-Xent loss [38]. For a positive pair of latent vectors (z_i, z_j), this loss is formulated as:

$$L_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

where $\text{sim}(u, v)$ is the cosine similarity, τ is a temperature hyperparameter, and $\mathbb{1}_{[k\neq i]}$ is an indicator function.Here, N represents the batch size of original samples. The summation in the denominator iterates up to $2N$ because each of the N samples in the batch is augmented twice (as described above), resulting in a total of $2N$ augmented representations. For any given representation z_i, one representation is its positive pair, and the remaining $2N - 2$ representations serve as negative pairs.

The encoder is trained for 100 epochs using an AdamW optimizer [44] with a learning rate of 3×10^{-4} and a weight decay of 7×10^{-4}. A Cosine Annealing scheduler [45] is employed to manage the learning rate. Minimizing this objective yields a powerful foundational encoder, f_θ, for subsequent stages.

3) Enabling Cross-Dataset Pre-training via the Unified Channel Schema: Pre-training on diverse EEG datasets is hindered by their inherent heterogeneity in channel counts and layouts (e.g., 62 in SEED [2], 32 in DEAP [3]). To overcome this barrier, we introduce the Unified Channel Schema (UCS), a framework enabling a robust univariate pre-training

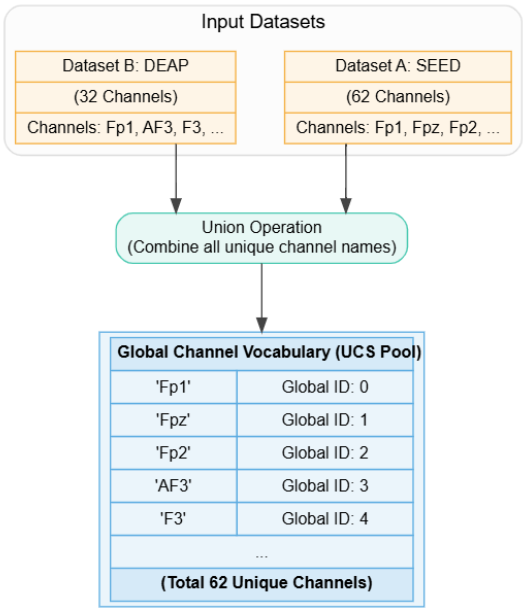


Fig. 2: Illustration of the Unified Channel Schema (UCS). Channels from heterogeneous datasets (e.g., SEED-62ch, DEAP-32ch) are mapped via a 'union' operation. **Notably, as all 32 channels from DEAP are a subset of the 62 channels from SEED,** this results in a global vocabulary containing all unique channels (62 total). Each channel name is assigned a Global ID, which is then used by the model's embedding layer to retrieve a learnable channel-specific representation.

paradigm that maximizes data utilization. Unlike baseline methods that discard data by using only the intersection of channels, the UCS creates a canonical Channel Symbol Pool from the union of all unique electrode identifiers across datasets. This pool serves as a global vocabulary for channel identity. During pre-training, each channel's local name is mapped to a global ID, which retrieves a corresponding learnable channel embedding. This embedding is then added to each temporal token's representation, infusing channel-specific information prior to the self-attention layers. Crucially, our pre-training encoder treats channels as discrete symbols, deliberately deferring the modeling of their spatial topology. This strategic decoupling allows the encoder to first learn universal temporal features from the maximal available data. The complex inter-channel spatial dependencies are then explicitly captured by a Graph Neural Network (GNN) [46] during the task-specific multivariate fine-tuning stage (see Section III-D). The UCS thus resolves structural misalignment across datasets, enabling the creation of a comprehensive foundational encoder at a scale previously unachievable.

An illustration of this process is provided in Figure 2.

C. Stage Two: Per-Subject Supervised Fine-tuning

While the foundational encoder developed in Stage One captures universal, subject-agnostic EEG features, it does not by itself address the critical challenge of 'high inter-subject variability'. To bridge this gap and specialize the universal model to each individual's unique neurophysiological characteristics, our framework proceeds directly to a single, per-subject supervised fine-tuning stage.

For each subject, we initialize the downstream classification architecture (detailed in Sec. III-D), which integrates the foundational **pre-trained encoder** (f_θ) as its core feature extractor. The entire network, including the pre-trained encoder and the randomly initialized classifier components (e.g., GAT and Transformer layers), is then fine-tuned on that subject's labeled training data for 100 epochs.

A critical component of this stage is the use of a **differential learning rate** [47] to prevent catastrophic forgetting of the rich representations learned during universal pre-training. As implemented in our code, we apply a significantly lower learning rate of 5×10^{-6} (e.g., 'lr_encoder') to the **pre-trained encoder**, while the newly added downstream components are trained with a higher learning rate of 1×10^{-4} (e.g., 'lr_head'). This strategy, managed by an AdamW optimizer [44], preserves the powerful pre-trained features while allowing the new classifier layers to learn the task-specific mapping effectively.

Notably, our final framework **omits** an intermediate, per-subject *self-supervised adaptation* stage. As demonstrated in our ablation study, this intermediate stage was not universally beneficial and proved detrimental on noisier datasets (e.g., DEAP). Therefore, our proposed framework converges on a robust and efficient two-stage paradigm: (1) Universal Pre-training and (2) Direct Supervised Fine-tuning.

D. Overall Model Architecture

Our framework is instantiated through two core architectural modules: our novel **ART (Adaptive Resampling Transformer)** encoder for self-supervised feature extraction and a hybrid **GAT-Transformer** network for multivariate classification.

1) Univariate Temporal Encoder (ART): The primary function of the univariate encoder is to perform deep temporal

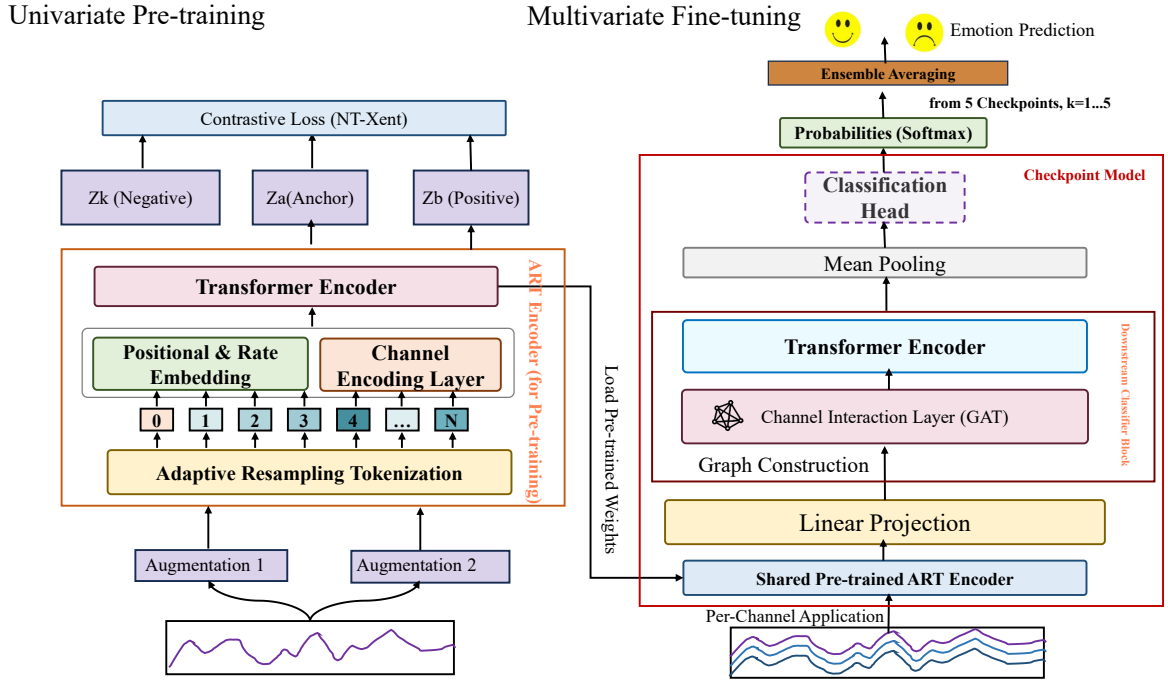


Fig. 3: The overall architecture of our proposed 'Univariate Pre-training + Multivariate Fine-tuning' framework. **(Left)** The **Univariate Pre-training** stage learns a shared Transformer Encoder (f_θ) using self-supervised contrastive learning on single, augmented EEG channels. **(Right)** The **Multivariate Fine-tuning** stage loads these weights (indicated by the 'Load Pre-trained Weights' arrow). This downstream classifier uses the shared **pre-trained encoder** (f_θ) as its core feature extractor. The encoder's weights are **fine-tuned** (typically with a low differential learning rate) as part of the larger spatio-temporal model, which then uses a Linear Projection, a Transformer Encoder, and a Channel Interaction Layer **(GAT)** to capture complex dependencies and make the final emotion prediction.

feature extraction on a per-channel basis. This function is performed by our novel **ART (Adaptive Resampling Transformer)** encoder.

The ART replaces the standard, fixed-patch input layer of a traditional Transformer with a sophisticated adaptive tokenizer. This design overcomes a critical flaw of fixed-patching, where a 16-sample patch captures a different temporal duration at different sampling rates (e.g., 125ms at 128Hz vs. 80ms at 200Hz), leading to semantic inconsistency.

Our ART architecture is composed of the following layers:

- **Adaptive Resampling Tokenization:** To handle heterogeneous sampling rates at the architectural level, our novel tokenizer module uses a two-stage process inspired by modern vision transformers. First, a stack of **1D Convolutional layers** acts as a feature extractor, scanning the entire raw EEG channel sequence (e.g., 256 samples at 128Hz or 400 samples at 200Hz) to learn local temporal patterns. Second, an **Adaptive Average Pooling** layer ('AdaptiveAvgPool1d') is applied to this variable-length feature map, which intelligently downsamples it into a **fixed sequence of 16 patches**. This ensures a uniform sequence length for the Transformer while guaranteeing that each patch is derived from the full context of the 2-second segment, regardless of the original signal's sampling rate. These 16 patches are then projected into the 128-dimensional latent space (d_{model}).
- **Contextual Embedding:** To retain full context, the token embeddings are augmented with three separate learnable encodings. The final input token sequence x_{in} is generated by the element-wise sum of these four distinct components:

$$x_{\text{in}} = x_{\text{patch}} + E_{\text{pos}} + E_{\text{chan}} + E_{\text{rate}}$$

Where each component serves a unique and critical purpose before being fed to the Transformer:

  - x_{patch}: This is the **content embedding** (with shape $\mathbb{R}^{16 \times d_{\text{model}}}$). This is the sequence of 16 patches derived from the raw signal by the ART tokenizer. It represents **'what'** is happening in the signal's content.
  - E_{pos}: This is the **positional encoding** (with shape $\mathbb{R}^{16 \times d_{\text{model}}}$). This is a standard learnable embedding [24] that is added to x_{patch}. Its job is to inform the Transformer **'where'** each patch is located in the sequence (e.g., distinguishing the 1st patch from the 16th).
  - E_{chan}: This is the **channel identity embedding** (with shape $\mathbb{R}^{d_{\text{model}}}$). A single vector is retrieved from an embedding table (of size 62) corresponding to the channel's identity (e.g., Fp1 vs. O2). This vector is then *broadcast* (copied) across all 16 patches and added, informing the model **'which channel'** this entire 16-patch sequence belongs to. This is critical

for enabling the encoder to learn channel-specific patterns.
  - E_{rate}: This is the novel **rate identity embedding** (with shape $\mathbb{R}^{d_{\text{model}}}$). Similar to E_{chan}, a single vector is retrieved from an embedding table (e.g., of size 2) corresponding to the signal's original sampling rate (e.g., 128Hz vs. 200Hz). This vector is also *broadcast* across all 16 patches, informing the model of the **'temporal context'** (i.e., how the 16 patches were derived).

  This mechanism, analogous to the input embedding process in models like BERT [6], effectively fuses content (x_{patch}), sequential order (E_{pos}), spatial identity (E_{chan}), and temporal identity (E_{rate}) into a single, rich representation (x_{in}) for the subsequent Transformer blocks.
- **Transformer Blocks:** The resulting sequence, x_{in}, is processed by a stack of **three** Transformer encoder layers (e_{layers}) [24]. Each layer utilizes an **8-head** multi-head self-attention mechanism (n_{heads}) and a position-wise feed-forward network with a hidden dimensionality of 256 (d_{ff}). This encoder is the sole component updated during the self-supervised pre-training stage. Crucially, because the ART tokenizer standardizes all inputs to a $(16 \times d_{\text{model}})$ sequence, this core Transformer architecture remains consistent and stable across all heterogeneous data sources.

2) Multivariate Spatio-Temporal Classifier: The multivariate classifier integrates the pre-trained encoders to perform the final emotion recognition task by explicitly modeling both dynamic and global spatio-temporal dependencies. The data processing pipeline is as follows:

- **Channel-wise Feature Extraction and Projection:** Given a multi-channel EEG sample with C channels, each channel is processed independently by a shared, pre-trained **ART** encoder (Sec. III-D1) to yield a set of C channel-specific feature vectors of 128 dimensions (the d_{model} of the encoder). These vectors are immediately passed through a linear projection layer, mapping them from 128 to the 256 dimensions (d_{model}) required by the subsequent modules.
- **Graph Attention-based Spatial Aggregation (GAT):** To model the inter-channel relationships, we replace the static GCN with a dynamic **Graph Attention Network (GAT)** [48]. We continue to employ a fully-connected graph structure where the C projected channel representations are conceptualized as nodes. The node features are then refined via **two layers of GAT** . Crucially, unlike the static, fixed-edge-weight GCN, the GAT layers use a multi-head self-attention mechanism (with **4 attention heads**) to learn dynamic, data-driven weights for each edge (channel-pair). This design enables the model to learn the functional connectivity between brain regions, 'attending' the most informative interactions for a given cognitive task rather than relying on static, pre-defined graph structures. This data-driven approach to relational modeling—a key advantage of GAT over GCN—is validated by our ablation study as essential for achieving state-of-the-art performance on high-noise datasets like DEAP. The attention mechanism of the GAT layer [48] is computed as follows. First, a shared linear transformation $\mathbf{W}$ is applied to every node (channel) embedding $\mathbf{h}_i$. Then, the attention coefficient e_{ij} between any two nodes i and j is calculated:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i || \mathbf{W}\mathbf{h}_j]\right) \tag{1}$$

where $||$ denotes concatenation and $\mathbf{a}$ is a learnable weight vector. These coefficients are then normalized across all of node i's neighbors $j \in \mathcal{N}_i$ using the softmax function to obtain the final attention weights α_{ij}:

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})} \tag{2}$$

The final output embedding for the node $\mathbf{h}'_i$ is then computed as the aggregated, weighted sum of its neighbors' features, followed by a non-linear activation σ (such as ELU):

$$\mathbf{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\mathbf{h}_j\right) \tag{3}$$

As our model utilizes **4 attention heads** (as mentioned previously), this entire process (from Eq. 1 to Eq. 3) is executed in parallel by $K = 4$ independent heads. The resulting 4 output embeddings are then concatenated to form the final, high-dimensional output representation for the node, which is then passed to the next layer.
- **Global Dependency Modeling and Prediction:** The spatially-aware channel representations from the GAT, $\mathbf{H}_{\text{gat}} \in \mathbb{R}^{C \times d_{\text{model}}}$, are subsequently fed into a **2-layer** Transformer encoder (e_{layers}) [24] with **4 attention heads** (n_{head}). This Transformer acts on the channel dimension (C), capturing global, long-range dependencies between all channel representations.
Each encoder layer applies Multi-Head Attention (MHA) and a Feed-Forward Network (FFN):

$$\mathbf{H}' = \text{LayerNorm}(\mathbf{H}_{\text{gat}} + \text{MHA}(\mathbf{H}_{\text{gat}})) \tag{4}$$

$$\mathbf{H}_{\text{out}} = \text{LayerNorm}(\mathbf{H}' + \text{FFN}(\mathbf{H}')) \tag{5}$$

To derive a single feature vector for the entire EEG segment, **mean pooling** is applied across the channel dimension (C) of the Transformer's output sequence, $\mathbf{H}_{\text{out}}$:

$$\mathbf{h}_{\text{pooled}} = \frac{1}{C}\sum_{i=1}^{C} \mathbf{H}_{\text{out}}[i, :] \tag{6}$$

Finally, this pooled representation $\mathbf{h}_{\text{pooled}}$ is passed to a linear classification head (with a dropout rate of 0.5) to produce the final emotion prediction logits, $\mathbf{z}$:

$$\mathbf{z} = \text{Linear}(\text{Dropout}(\mathbf{h}_{\text{pooled}})) \tag{7}$$

The final probabilities are obtained by applying a softmax function to the logits $\mathbf{z}$ during inference.

E. Final Evaluation via Checkpoint Ensembling

To enhance predictive accuracy and mitigate the variance associated with a single training run, we employ a checkpoint ensembling strategy for the final evaluation. This technique leverages multiple high-performing model states discovered during the optimization process at no additional training overhead.

The procedure is integrated into the fine-tuning stage (Sec. III-C). Instead of selecting only the model with the single best validation score, we maintain a collection of the **Top-5** model checkpoints that achieve the highest accuracy on the validation set. These checkpoints are not necessarily from the final epochs but can be from any point where a new peak in validation accuracy was observed.

During inference, these five models form an ensemble. For any given input x, each model M_k ($k \in \{1, \ldots, 5\}$) independently generates a posterior probability distribution. These are then averaged to produce a final, ensembled probability distribution, $P_{\text{ensemble}}(y|x)$, as follows:

$$P_{\text{ensemble}}(y|x) = \frac{1}{5} \sum_{k=1}^{5} \text{softmax}(M_k(x)) \tag{8}$$

where $M_k(x)$ represents the output logits from the k-th model. The final predicted class, $\hat{y}$, is the one with the highest probability:

$$\hat{y} = \underset{c}{\text{argmax}}\, P_{\text{ensemble}}(y = c|x) \tag{9}$$

The rationale for this approach is twofold. First, the stochastic nature of training means that different epochs may yield models specialized in slightly different, yet effective, feature subspaces. Ensembling these provides a more comprehensive view of the solution landscape. Second, this model averaging is a well-established technique for reducing prediction variance and improving generalization, as it smooths out the decision boundary and is less sensitive to the idiosyncrasies of any single checkpoint.

IV. Experiments

A. Experiment Setting

a) Datasets: We validate our framework on three widely-used public EEG emotion recognition datasets: SEED [2], DEAP [3], and DREAMER [4]. These datasets are evaluated on different emotion classification tasks: SEED involves 3-class classification (Positive, Neutral, Negative), while DEAP and DREAMER are evaluated on binary valence classification (High/Low). A key challenge, which motivates our work, is the profound heterogeneity of these datasets. This manifests in two primary ways: (1) differences in channel configurations (62 channels for SEED, 32 for DEAP, and 14 for DREAMER), and (2) variations in sampling rates.**To address temporal heterogeneity, we move away from destructive pre-processing (i.e., uniform resampling).** Instead, our framework is designed to handle varying sampling rates (e.g., 128 Hz, 200 Hz) directly at the architectural level using our novel **ART (Adaptive Resampling Transformer)** encoder (detailed in Sec. III-D1). Subsequently, our proposed Unified Channel Schema (UCS), as detailed in Section III-B3, is designed to address the challenge of channel heterogeneity and enable joint pre-training across these disparate data sources.

b) Data Preprocessing using MNE-Python: To ensure consistency and mitigate artifacts across the heterogeneous datasets (SEED, DEAP, DREAMER), a standardized preprocessing pipeline was implemented using the MNE-Python library [49]. First, raw EEG data were loaded from their original formats (e.g., '.mat' for SEED, pickled '.dat' for DEAP) into MNE 'Raw' objects. Appropriate channel location information (montages) was assigned using standard templates ('biosemi32' for DEAP) or custom configurations (SEED 10-20 system) via MNE's channel management functions. Following the common practice referenced in the literature, an average reference was applied. Subsequently, band-pass filtering between 1 Hz and 49 Hz was applied to remove baseline drift and high-frequency noise. Artifact removal was performed using Independent Component Analysis (ICA). We employed MNE's ICA implementation, fitting it to the filtered data. Critically, artifactual components (e.g., eye blinks, muscle activity) were automatically identified and labeled using the 'mne_icalabel' library, which integrates the ICLabel model [50]. Components labeled as non-brain sources were automatically excluded. Finally, the cleaned continuous data were segmented into 2-second epochs with a 0.2-second overlap. These processed epochs were saved in the MNE-native '.fif' format for subsequent model training and evaluation.

c) Evaluation Protocol: Our primary evaluation protocol follows the standard **within-subject** approach [1]. For each subject in a given dataset, their data is partitioned into a training set (80%) and a test set (20%). A personalized model is then trained *exclusively* on that subject's training set and evaluated on their corresponding test set. This protocol, distinct from cross-subject evaluation (which typically tests generalization to unseen subjects without subject-specific fine-tuning) [5], is specifically chosen here to rigorously assess the framework's ability to leverage the universal pre-trained features and effectively adapt to create high-performance, **personalized classifiers for each individual** using only their own data. The final reported result is the average accuracy across all subjects in the dataset.

d) Evaluation Metrics: The primary metric used for performance evaluation is classification **Accuracy** (%). We report the mean accuracy and standard deviation across all subjects for each experiment.

e) Baseline Settings: To comprehensively evaluate our proposed framework, we compare it against two types of baselines:

- **Internal Baseline ('From Scratch')**: To validate the efficacy of our universal pre-training,we compare our full model against an identical architecture trained without leveraging any of the self-supervised pre-training stages. This 'From Scratch' baseline is initialized randomly (as is standard practice) and trained only on the downstream task's labeled data, serving to isolate and quantify the direct benefits of our pre-training paradigm.
- **External SOTA Baselines**: We compare our model's within-subject performance against recent state-of-the-

art methods. For fair comparison, we categorize external SOTA baselines into two distinct approaches: (1) end-to-end models, which learn features directly from raw data, and (2) methods based on extensive feature engineering, which utilize handcrafted features.

f) Implementation Details: All experiments were conducted on a platform equipped with an NVIDIA A100-SXM4-40GB GPU, using Python 3.9.20, PyTorch [51], and MNE-Python [49]. For reproducibility, the random seed was fixed to 42.

g) Hyperparameter Settings: Our framework involves a two-stage paradigm training process with distinct hyperparameter settings:

- **Universal Pre-training Stage**: The universal encoder was trained for a maximum of 100 epochs with a batch size of 512. We used the AdamW optimizer [44] with a learning rate of 3×10^{-4} and a weight decay of 7×10^{-4}, along with a Cosine Annealing scheduler [45].
- **Supervised Fine-tuning Stage**: During the final subject-dependent fine-tuning, a differential learning rate was used: 5×10^{-6} for the pre-trained encoder and 1×10^{-4} for the new classifier layers. The model was trained for up to 100 epochs with a batch size of 128.

B. Main Results

1) Universal Pre-training Enables Cross-Dataset Knowledge Transfer: A fundamental challenge in creating a universal pre-training model is handling the channel heterogeneity across datasets. To address this, we evaluated two strategies on the DREAMER dataset after pre-training on a mixture of SEED and DEAP: a baseline approach using the **intersection** of channels, and our proposed **Unified Channel Schema (UCS)** which utilizes the **union** of all available channels. The results, presented in Table I, provide several key insights. To provide an ideal benchmark, we also computed an 'Ideal Upper Bound' by performing within-domain pre-training and fine-tuning using only the DREAMER dataset (using our new ART+GAT architecture).

First, our new **ART+GAT** architecture demonstrates vastly superior knowledge transfer capabilities. Our intersection-based model achieved **94.08%** accuracy, while our UCS-based model achieved **93.05%**.

Second, and most remarkably, our cross-dataset transfer model (using Intersection) at **94.08% surpassed the Ìdeal Upper Bound'** (93.93% from within-domain pre-training). This provides powerful evidence that our universal pre-training captures generalized knowledge that is even more effective than representations learned from only the target domain's data.

Third, pre-training provides clear and essential gains over training our ART+GAT model from scratch (which achieved a strong **90.38%** baseline). The Intersection strategy yielded a **+3.70%** absolute improvement, while UCS yielded a **+2.67%** improvement, confirming the value of universal pre-training.

Finally, while the simpler intersection method achieved the highest accuracy in this specific transfer task, the UCS framework (93.05%) remains highly competitive. The true value of UCS lies in its superior scalability and future-proof design. The intersection method is not applicable if future datasets have few or no channels in common. In contrast, the UCS provides a robust and scalable paradigm capable of integrating any number of heterogeneous datasets, making it the only viable path towards truly large-scale EEG foundation models.

TABLE I: Impact of Different Pre-training Strategies on DREAMER (Within-Subject)

Configuration	Description	Acc. (%)
Baseline	ART+GAT (Training from Scratch)	**90.38**
Cross-Dataset Pre-training (on SEED+DEAP)		
Intersection	ART+GAT using channel intersection	**94.08**
UCS	ART+GAT using channel union (UCS)	**93.05**
Ideal Upper Bound	ART+GAT (Within-Domain Pre-training on DREAMER)	**93.93**

2) State-of-the-Art Performance in Personalized Emotion Recognition: We further evaluated our complete framework against recent state-of-the-art (SOTA) methods on the standard within-subject benchmark. The results, presented in Table II, demonstrate the clear superiority of our proposed **ART+GAT** framework. Our model achieves new state-of-the-art performance across all three benchmark datasets.

- On the **SEED** dataset, our model achieves a near-perfect accuracy of **99.27%**, surpassing previous SOTA methods.
- Most notably, on the highly challenging and noisy **DEAP** dataset, our model achieves a breakthrough accuracy of **93.69%**. This SOTA performance is directly attributable to our architectural choice, validating the critical role of our **GAT** module in handling high-noise signals. As demonstrated in our ablation study (Sec. IV-C2), this module provides a massive **+22.19%** absolute improvement where simpler GCNs fail.
- On the **DREAMER** dataset, our model also achieves a new SOTA performance of **93.93%**.

These results confirm that our end-to-end paradigm, combining the **ART** encoder's adaptive temporal feature extraction with the **GAT**'s robust spatial attention mechanism, establishes a new and significantly higher benchmark for personalized EEG emotion recognition.

C. Ablation Study and Model Interpretation

1) The Efficacy of Pre-training - A Necessary Stabilizer: To fundamentally validate the contribution of our universal pre-training paradigm, we compared the performance of our best models (ART+GAT) against an identical architecture trained 'from scratch'. As summarized in Table III, the results provide compelling evidence that pre-training is not merely a 'performance booster' but a crucial **'stabilizer' and 'necessity'** for achieving robust SOTA performance. Our analysis revealed two key findings:

- **Pre-training prevents convergence to sub-optimal solutions.** On the challenging **DEAP** dataset, the 'from scratch' model converged to a much lower accuracy

TABLE II: Within-Subject SOTA Performance Comparison on SEED, DEAP, and DREAMER

Dataset	Model	Methodology	Accuracy (%)
SEED	**Our Model**	**End-to-End (ART + GAT)**	**99.27**
	DenseNet-based [22]	Feature Engineering (DE)	96.73
	SI-CLEER [42]	End-to-End	95.45
	DGAT [52]	Feature Engineering (DE)	94.00
	EEG-DisGCMAE [43]	Feature Engineering (DE)	93.60
	CMHFE [53]	End-to-End	82.81
DEAP*	**Our Model**	**End-to-End (ART + GAT)**	**93.69**
	DGAT [52]	Dynamic Graph Attention Net	93.55
	2D-CNN-LSTM [54]	CNN + LSTM	91.92
	CNN-KAN [55]	CNN + KAN (Sparse)	90.03
	Wu et al. (2025) [56]	EEG + EOG Fusion	86.61
	ERTNet [57]	End-to-End	73.31
DREAMER*	**Our Model**	**End-to-End (ART + GAT)**	**93.93**
	ELSTNN [58]	Spatiotemporal NN	93.72
	Two-stream net [59]	Hierarchical CNN	93.01
	CMHFE [53]	End-to-End	87.12
	DGCNN [23]	Dynamical GCNN	86.23

* Accuracy reported for the binary Valence classification task.

(86.04%), while the pre-trained model achieved our SOTA result (93.69%). This **+7.65%** absolute gain demonstrates that pre-training is essential for guiding the complex ART+GAT model out of local minima towards the global optimum. A similar significant gain of **+3.55%** was observed on **DREAMER** (93.93% vs. 90.38%).

- **Pre-training prevents catastrophic training collapse and finds a superior optimum.** Most critically, pre-training acts as an essential stabilizer. While our pre-trained model achieved a stable SOTA performance of **99.27%** on **SEED**, the 'from scratch' model exhibited severe instability. In a more rigorous test on the final subject, we found that **2 out of 3 attempts** with different random seeds resulted in a **complete training collapse**. Furthermore, even the single successful attempt, which converged with a third seed, only achieved an accuracy of **97.96%**. This result, while successful, remains significantly lower than our pre-trained SOTA (99.27%), proving it is a sub-optimal solution. This case study provides definitive proof that our pre-training paradigm is a prerequisite for both **ensuring reliable convergence** (avoiding collapse) and **achieving the true SOTA performance** (avoiding sub-optimal solutions).

This evidence demonstrates that our self-supervised pre-training learns highly meaningful and transferable representations that are essential for navigating the complex optimization landscape of deep EEG models. The t-SNE visualizations in Figure III further corroborate this, showing that the pre-trained model organizes features into much clearer subject-specific clusters compared to the chaotically mixed features from the model trained from scratch.

2) Analysis of Architectural Components: Having established the necessity of pre-training, we performed a series of in-depth ablation studies to validate each component of our final SOTA **ART+GAT** architecture. The results, presented in Table IV, are analyzed below based on our 'Three Pillars' framework.

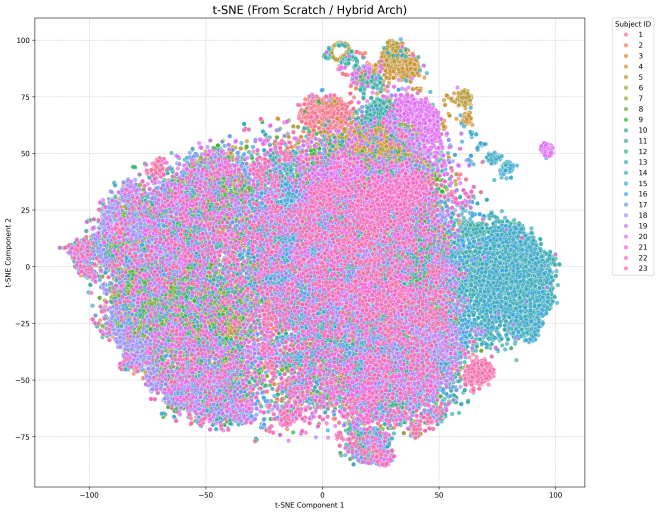
TABLE III: Efficacy of Universal Pre-training (ART+GAT vs. From Scratch)

Dataset	**From Scratch (%)**	**With Pre-training (%)**	**Gain (%)**
DEAP	86.04	**93.69**	**+7.65**
DREAMER	90.38	**93.93**	**+3.55**
SEED	97.96 (Unstable)	**99.27**	**+1.31 (Stable)**

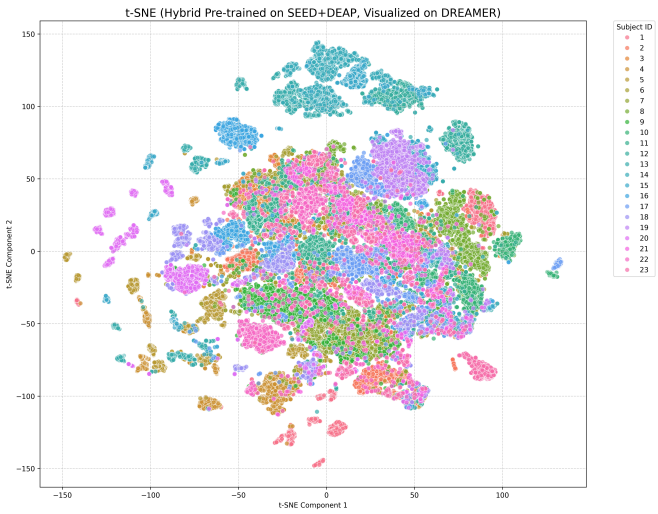
a) Pillar I: The Necessity of Graph Networks (GAT vs. NoGraph): We first tested the hypothesis that a multi-variable graph network is essential. We compared our full model against a variant where the GAT module was removed entirely (**ART + NoGraph**). The results, now confirmed across all three datasets, were definitive. On **DEAP**, this caused a catastrophic performance drop of **-16.44%** (from 93.69% to 77.25%). On **SEED**, the drop was also highly significant at **-5.05%** (from 99.27% to 94.22%), and a similar significant drop of **-3.82%** was observed on **DREAMER** (from 93.93% to 90.11%). This irrefutably proves that the graph network is an indispensable component for modeling inter-channel relationships on all evaluated datasets.

b) Pillar II: The Superiority of GAT (GAT vs. GCN): Next, we validated our choice of GAT over a standard GCN, as the GCN was used in the previous framework. This was the most critical factor on high-noise datasets. On **DEAP**, replacing GAT with a GCN (**ART+ GCN**) caused the model's performance to collapse from 93.69% to 71.50%. This massive **+22.19%** advantage for GAT demonstrates that its dynamic attention mechanism is the key to handling signal noise and unlocking SOTA performance, whereas the GCN's static weights are insufficient.

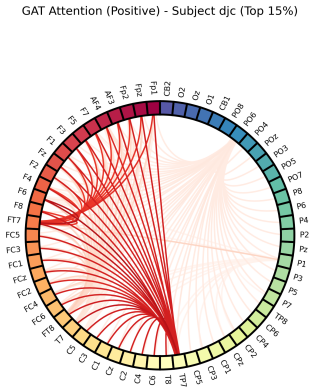
c) Pillar III: The Advantage of the ART Encoder (ART vs.Fixed-Patch Enc): Finally, we validated the contribution of our novel **ART** temporal encoder itself. We compared our SOTA model (ART + GAT) against a baseline where the ART module (with its adaptive tokenizer) was replaced by



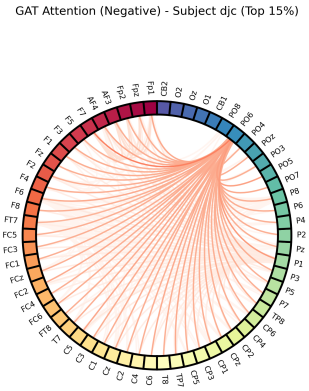
(a) From Scratch



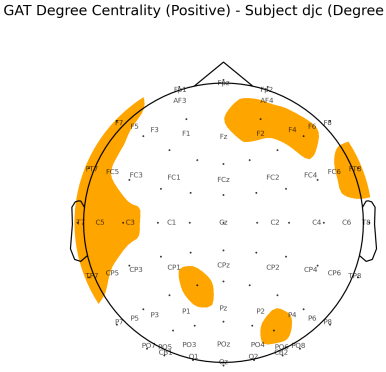
(b) With Universal Pre-training



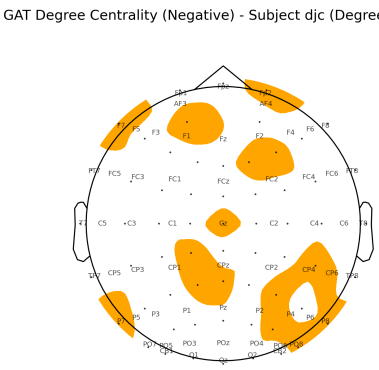
(c) GAT attention connectivity (Top 15%) for 'Positive' emotion (Subject 'djc').



(d) GAT attention connectivity (Top 15%) for 'Negative' emotion (Subject 'djc').



(e) Key channels from GAT degree centrality (Degree $\geq$ 8) for 'Positive' emotion.



(f) Key channels from GAT degree centrality (Degree $\geq$ 8) for 'Negative' emotion.

Fig. 4: Universal pre-training learns semantically structured representations, and the GAT module captures emotion-specific neural patterns. (Top Row) t-SNE visualization of feature embeddings from the DREAMER dataset. (a) The model trained from scratch produces a largely undifferentiated feature space. (b) In stark contrast, our pre-trained model transforms this space, organizing the features into highly structured manifolds. (Middle and Bottom Rows) GAT interpretability analysis for a single representative subject ('djc') comparing distinct emotional states. (Middle Row) (c) and (d) visualize the top 15% of GAT attention weights, revealing different functional connectivity patterns for Positive vs. Negative emotions. (Bottom Row) (e) and (f) highlight the most critical channels (nodes) identified via degree centrality (Degree $\geq$ 8), showing a clear shift in the model's spatial focus (e.g., from left-temporal (e) to bilateral posterior-parietal (f)) depending on the emotion being processed.

TABLE IV: Ablation Study of SOTA (ART+GAT) Architectural Components

Config.	Model Configuration	SEED (%)	DEAP (%)	DREAMER (%)
(A)	**Our SOTA Model (ART + GAT)**	**99.27**	**93.69**	**93.93**
	— Architectural Ablations (Pillars) —			
(B)	*Pillar I:* No Graph (ART + NoGraph)	94.22	77.25	90.11
(C)	*Pillar II:* Use GCN (ART + GCN)	98.24	71.50	93.46
(D)	*Pillar III:* Use Fixed-Patch Enc. (Fixed-Patch Enc. + GAT)	98.86	93.61	93.13

a standard Transformer using a simpler **fixed-patch tokenization** strategy (termed **'Fixed-Patch Enc.' + GAT**). Our ART encoder demonstrated consistent, superior performance across all datasets: **SEED (+0.41%)** (99.27% vs. 98.86%), **DEAP (+0.08%)** (93.69% vs. 93.61%), and **DREAMER (+0.80%)** (93.93% vs. 93.13%). This confirms the advantage of our adaptive tokenizer over a standard, fixed-patch approach.

3) Model Interpretability: Visualizing Personalized Attention: To explore the model's decision-making process, we visualized the attention weights from the **GAT module** (see Section III-D2) to understand its learned spatial focus during classification. For this case study, we selected a representative subject ('djc') from the SEED dataset.

We investigated the *independent* attention patterns for 'Positive' and 'Negative' emotional states separately, rather than a direct contrastive map. This two-part analysis, inspired by prior work (as shown in your provided reference), is presented in Figure 4 (Middle and Bottom Rows).

- **Connectivity Patterns (Fig. 4c, 4d):** The middle row visualizes the functional connectivity by plotting the top 15% of GAT attention weights (a threshold determined by our analysis script). This reveals distinct network structures for each emotion. For this subject, the 'Positive' state (Fig. 4c) is characterized by strong, focused connections, particularly in the left hemisphere. In contrast, the 'Negative' state (Fig. 4d) exhibits a more diffuse, bilateral connectivity pattern.
- **Key Channel Identification (Fig. 4e, 4f):** The bottom row illustrates a *degree centrality* analysis to identify the most critical channels (nodes) in these networks. Following our analysis script, key channels are defined as those with a connection degree ≥ 8. The resulting topographic maps show a clear spatial shift in the model's focus. For 'Positive' emotion (Fig. 4e), the model's focus is concentrated in the left-temporal and parietal regions. For 'Negative' emotion (Fig. 4f), this focus shifts significantly to bilateral posterior-parietal and central regions.

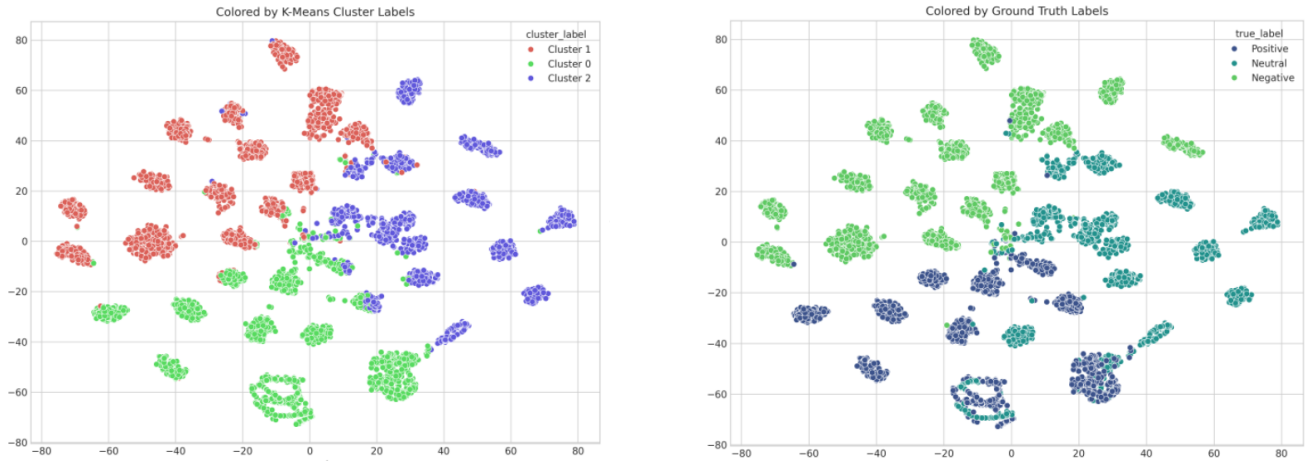
This analysis provides clear visual evidence that our personalized fine-tuning framework captures nuanced, emotion-specific neural patterns at the individual level. Instead of enforcing a single, flawed universal rule, the personalized GAT model learns to dynamically shift its spatial attention—focusing on different functional networks and key brain regions—to distinguish between cognitive states for that specific subject. This demonstrates the model's interpretability and further validates the necessity of our personalized approach.

4) Visualization of the Learned Feature Space: To further complement our quantitative results and the spatial attention analysis, we investigated the quality of the learned representations at the individual subject level after the complete fine-tuning process. We projected the final layer's feature embeddings from a representative subject from the SEED dataset ('djc') into a two-dimensional space using the t-SNE algorithm [60] to visualize their structure. The results are presented in Figure 5.

Figure 5 provides a compelling qualitative validation of our framework's personalization capabilities. In Figure 5b, where the data points are colored by their ground-truth labels, we observe the formation of three distinct and well-separated clusters, corresponding directly to the 'Positive', 'Neutral' and 'Negative' emotional states. This clear separation in the embedding space demonstrates that the fine-tuned model has successfully learned a highly discriminative feature manifold, which is essential for the high classification accuracies reported in Section IV-B. To further probe the intrinsic structure of this learned space, we applied an unsupervised K-Means clustering algorithm [61] with $k = 3$ to the same set of embeddings. Figure 5a displays the results, with points colored by their assigned cluster ID. A striking correspondence is evident between the algorithmically discovered clusters and the ground-truth emotional categories. This alignment signifies that the primary axes of variation captured by our model are not merely correlational but are semantically meaningful, mapping directly onto the core emotional dimensions of the task. This provides strong evidence that our universal pre-training, followed by subject-specific fine-tuning, yields not just a performant classifier, but also a robust and well-structured feature representation at a granular, per-subject level. Complementing the cross-subject visualization in Figure 4b, which shows that pre-training effectively separates different subjects, Figure 5 demonstrates that the subsequent fine-tuning successfully organizes the feature space for a single subject into semantically meaningful emotional clusters. This two-stage process—first separating subjects, then separating their emotions—highlights the comprehensive learning capability of our framework.

V. Conclusion

In this work, we addressed the critical challenge of data heterogeneity in EEG-based emotion recognition, which hinders the development of generalizable deep learning models. We proposed 'One Model for All', a novel universal pre-training framework designed to effectively leverage data across disparate EEG datasets. Our core contribution is a 'decoupled learning paradigm' featuring 'univariate pre-training' via contrastive self-supervised learning, enabled by the 'Unified Channel Schema (UCS)' to handle channel heterogeneity. This pre-trained encoder, our novel **ART (Adaptive Resampling Transformer)**, then serves as a powerful foundation. We



(a) Colored by K-Means Cluster Labels

(b) Colored by Ground Truth Labels

Fig. 5: t-SNE visualization of the learned feature representations for a single subject after personalized fine-tuning. (a) Embeddings colored by cluster labels assigned by an unsupervised K-Means algorithm ($k = 3$). (b) The same embeddings colored by their ground-truth emotion labels. The clear formation of distinct clusters corresponding to the true labels and their strong alignment with the K-Means results validate the model's ability to learn a highly discriminative and semantically structured feature space.

demonstrated that this must be paired with a downstream **GAT (Graph Attention Network)** classifier, which is fine-tuned for subject-specific emotion recognition.

Our comprehensive experiments demonstrated the clear superiority of this new architecture. Our framework achieves new state-of-the-art (SOTA) performance across all three benchmark datasets: SEED (99.27%), DEAP (93.69%), and DREAMER (93.93%). The result on the high-noise DEAP dataset is particularly significant. Critically, we also showed SOTA knowledge transfer, with our model pre-trained on SEED+DEAP achieving **94.08%** on the unseen DREAMER dataset, a result that surpasses even the within-domain pre-trained 'ideal upper bound'.

Our in-depth ablation studies validated this architecture, proving that: (1) the pre-training stage is an essential **stabilizer** that prevents training collapse (e.g., a failure case on SEED); (2) the graph network is an **indispensable component** (a -16.44% drop on DEAP without it); and (3) the **GAT**'s attention mechanism is the critical key, providing a massive **+22.19% gain over a GCN** on the noisy DEAP dataset. Finally, interpretability analyses confirmed that our model learns neuro-plausible, highly individualized representations, validating the personalized fine-tuning approach.

Our work establishes a new state-of-the-art for personalized, **within-subject** emotion recognition, a critical goal for user-specific affective computing applications. The scalability of our universal pre-training framework, enabled by the **UCS** and the **ART** encoder, also opens up promising directions for future research. This pre-training methodology could be extended to encompass more datasets and diverse EEG paradigms (e.g., BCI, sleep staging), exploring its potential as a foundational model for various downstream EEG analysis tasks.

In conclusion, our proposed framework represents a significant step towards building universal, pre-trained models for EEG analysis. By effectively decoupling representation learning, introducing the **UCS** to address channel heterogeneity, and—most importantly—validating a new SOTA architecture composed of the **ART** encoder and the **GAT** classifier, we have demonstrated a scalable and powerful approach for leveraging diverse EEG data, paving the way for more robust and generalizable affective computing systems.

ACKNOWLEDGMENTS

This Work was supported by the Major Innovation Project for the Integration of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (2025ZDYS01, 2024GH24), the Jinan '20 New Colleges and Universities' Funded Project (202333043), and the Taishan Scholars Program (NO.tspd20240814).

REFERENCES

[1] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen, "Eeg based emotion recognition: A tutorial and review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 5, pp. 1–34, 2022. [Online]. Available: https://doi.org/10.1145/3524499

[2] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

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

[4] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," in *2017 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, 2017, pp. 1186–1191.

[5] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[8] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu, "Ts2vec: Towards universal representation of time series," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 8, 2022, pp. 8980–8987.

[9] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li, "Time-series representation learning via temporal and contextual contrasting," *arXiv preprint arXiv:2106.14112*, 2021.

[10] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Proceedings of the Machine Learning for Health NeurIPS Workshop*, vol. 136. PMLR, 2020, pp. 238–253.

[11] S.-H. Oh, Y.-R. Lee, and H.-N. Kim, "A novel eeg feature extraction method using hjorth parameter," *International Journal of Electronics and Electrical Engineering*, vol. 2, no. 2, pp. 106–110, 2014.

[12] V. Gupta, M. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using deep convolutional neural networks on wavelet transformed eeg," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 456–459.

[13] R.-Z. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for eeg-based emotion classification," *Tsinghua Science and Technology*, vol. 18, no. 5, pp. 449–459, 2013.

[14] J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela, "Measuring phase synchrony in brain signals," *Human brain mapping*, vol. 8, no. 4, pp. 194–208, 1999.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[16] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] R.-z. Li, X. Li, Y. Wang, and H.-w. Li, "Hierarchical convolutional neural networks for eeg-based emotion recognition," in *2018 International conference on artificial neural networks (ICANN)*. Springer, 2018, pp. 686–695.

[19] Z. Abbasvandi and A. Ghassemi, "Self-supervised learning for eeg-based emotion recognition," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 681–684.

[20] C. Du, W. Liu, T. Yang, R. Tao, and X. Yang, "An efficient lstm network for emotion recognition from multichannel eeg signals," *IEEE Access*, vol. 8, pp. 170 247–170 260, 2020.

[21] J. Chen, Y. Xin, and W.-D. Zhang, "Hierarchical gru network for eeg-based emotion recognition," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 1220–1224.

[22] N. Li, G.-w. Liu, J.-w. Yan, and C.-l. Li, "A novel emotion recognition method based on 1d-densenet," *Biomedical Signal Processing and Control*, vol. 87, p. 105342, 2024.

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

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[25] Y. Gong, J. Xu, B. Wang, W. Li, and T. Zhang, "Eeg-transformer: Self-attention based spatial-temporal feature extraction for eeg-based emotion recognition," *Computers in Biology and Medicine*, vol. 161, p. 107011, 2023.

[26] H. Jiang, D. Wu, Y. Wang, and B.-L. Lu, "Elastic graph neural network for eeg-based emotion recognition," *IEEE Transactions on Affective Computing*, 2023.

[27] Y. Wu, X. Liu, and J. Liu, "Fc-cnn-gru: A fusion model for eeg-based emotion recognition," *Computers in Biology and Medicine*, vol. 168, p. 107722, 2024.

[28] A. Apicella, P. Arpaia, and R. Maffettone, "Toward universal models for eeg-based emotion recognition: A cross-corpus study," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2024.

[29] M. Kukhilava, G. Gogniat, P. S. de Chazal, and H. P. K. J. Mac, "An evaluation framework for eeg-based emotion recognition models: A cross-dataset and cross-subject study," *Expert Systems with Applications*, vol. 261, p. 124844, 2025.

[30] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[31] J. Li, S. Qiu, C. Du, Y. Wang, and L. Zhang, "A novel adversarial domain adaptation network for eeg-based emotion recognition," *Sensors*, vol. 20, no. 7, p. 1964, 2020.

[32] A. Rafiei and J. K. Tsotsos, "Self-supervised learning for human activity recognition using physiological signals,"

Neural Networks, vol. 146, pp. 144–155, 2022.

[33] J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in *2018 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, 2018, pp. 415–419.

[34] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "EEG-GAN: Generative adversarial networks for electroencephalograhic (eeg) brain signals," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 2551–2555.

[35] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "Csdi: Conditional score-based diffusion models for probabilistic time series imputation," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 24 804–24 816.

[36] J. Dong, H. Yang, L. Chen, and Y. Yang, "Simmtm: A simple framework for masked time-series modeling," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 2023, pp. 2827–2840.

[37] Z. Hou, X. Liu, Y. Cen, Y. Dong, Z. Yang, and J. Tang, "Graphmae: Self-supervised masked graph autoencoders," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 594–604.

[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[39] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[40] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[41] J. Huang, Y. Song, D. Zhang, and C. Zhang, "Contrastive and generative self-supervised learning for human activity recognition," in *Proceedings of the 2023 ACM international joint conference on pervasive and ubiquitous computing & proceedings of the 2023 ACM international symposium on wearable computers*, 2023, pp. 31–36.

[42] X. Li, J. Song, Z. Zhao, C. Wang, D. Song, and B. Hu, "A supervised information enhanced multi-granularity contrastive learning framework for EEG based emotion recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 2325–2329.

[43] X. Wei, K. Zhao, Y. Jiao, H. Xie, L. He, and Y. Zhang, "Pre-training graph contrastive masked autoencoders are strong distillers for EEG," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, vol. 267. PMLR, 2025, arXiv:2411.19230.

[44] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[45] ——, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[46] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[47] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 328–339.

[48] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[49] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, no. 267, pp. 1–13, 2013.

[50] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "ICLabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, 2019.

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[52] Y. Ding, J. Yan, X. Li, Y. Chen, and J. Wang, "DGAT: a dynamic graph attention neural network framework for eeg emotion recognition," *Frontiers in Psychiatry*, vol. 16, p. 1633860, 2025.

[53] M. Hilali, A. Ezzati, and S. Ben Alla, "CMHFE-DAN: A transformer-based feature extractor with domain adaptation for EEG-based emotion recognition," *Information*, vol. 16, no. 7, p. 560, 2025.

[54] T. Wang, X. Huang, Z. Xiao, W. Cai, and Y. Tai, "Eeg emotion recognition based on differential entropy feature matrix through 2d-cnn-lstm network," *EURASIP Journal on Advances in Signal Processing*, vol. 2024, no. 1, p. 49, 2024.

[55] F. Xiong, M. Fan, X. Yang, C. Wang, and J. Zhou, "Research on emotion recognition using sparse eeg channels and cross-subject modeling based on cnn-kan-f 2 ca model," *PLoS One*, vol. 20, no. 5, p. e0322583, 2025.

[56] C. Li, S. Zhang, Y. Mu, L. Yang, Y. Peng, F. Li, Y. Zhang, Z. Liang, Z. Cao, F. Wan *et al.*, "Emotion recognition by learning the manifold of fused multiscale information of eeg signals," *IEEE Transactions on Affective Computing*, 2025.

[57] G. Liu, Y. Chen, X. Liu, and Y. Yin, "Ertnet: an interpretable transformer-based framework for eeg emotion recognition," *Frontiers in Neuroscience*, vol. 18, p. 1320645, 2024.

[58] D. Huang, D. Jiang, S. Zhoua, L. Chena, J. Linb, and E. Cambriac, "Emotional lateralization-inspired spatiotemporal neural networks for eeg-based emotion recognition," 2023.

[59] S. Gong, K. Xing, A. Cichocki, and J. Li, "Deep learning in eeg: Advance of the last ten-year critical period," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 348–365, 2021.

[60] L. Van der Maaten and G. Hinton, "Visualizing data using

t-sne," *Journal of machine learning research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[61] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.