



Article

EEG–fNIRS Cross-Subject Emotion Recognition Based on Attention Graph Isomorphism Network and Contrastive Learning

Bingzhen Yu, Xueying Zhang * and Guijun Chen

College of Electronic Information Engineering, Taiyuan University of Technology, Taiyuan 030024, China; ybzcasia@gmail.com (B.Y.); chenguijun@tyut.edu.cn (G.C.)

* Correspondence: zhangxy@tyut.edu.cn

Abstract

Background/Objectives: Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) can objectively capture the spatiotemporal dynamics of brain activity during affective cognition, and their combination is promising for improving emotion recognition. However, multi-modal cross-subject emotion recognition remains challenging due to heterogeneous signal characteristics that hinder effective fusion and substantial inter-subject variability that degrades generalization to unseen subjects. **Methods:** To address these issues, this paper proposes DC-AGIN, a dual-contrastive learning attention graph isomorphism network for EEG–fNIRS emotion recognition. DC-AGIN employs an attention-weighted AGIN encoder to adaptively emphasize informative brain-region topology while suppressing redundant connectivity noise. For cross-modal fusion, a cross-modal contrastive learning module projects EEG and fNIRS representations into a shared latent semantic space, promoting semantic alignment and complementarity across modalities. **Results:** To further enhance cross-subject generalization, a supervised contrastive learning mechanism is introduced to explicitly mitigate subject-specific identity information and encourage subject-invariant affective representations. Experiments on a self-collected dataset are conducted under both subject-dependent five-fold cross-validation and subject-independent leave-one-subject-out (LOSO) protocols. The proposed method achieves 96.98% accuracy in four-class classification in the subject-dependent setting and 62.56% under LOSO. Compared with existing models, DC-AGIN achieves SOTA performance. **Conclusions:** These results demonstrate that the work on attention aggregation, cross-modal and cross-subject contrastive learning enables more robust EEG–fNIRS emotion recognition, thus supporting the effectiveness of DC-AGIN in generalizable emotion representation learning.



Academic Editor: Estate (Tato) Sokhadze

Received: 30 December 2025

Revised: 23 January 2026

Accepted: 27 January 2026

Published: 28 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: EEG; fNIRS; emotion recognition; graph isomorphism network; contrastive learning

1. Introduction

Emotion is a core component of human cognition and decision making. It not only shapes individuals' subjective well-being and social interactions but also deeply participates in higher-order cognitive processes such as attention, memory, and value-based judgment [1]. With the rapid development of affective computing [2] and affective brain-computer interfaces [3], accurately recognizing an individual's affective state in complex real-world scenarios has become a fundamental problem for intelligent human–computer

interaction and mental-health monitoring. Traditional emotion recognition methods are often built on non-physiological cues—such as facial expressions, speech prosody, behavior, and text—which are explicitly observable but inherently subjective and vulnerable to intentional masking. In contrast, physiology-based emotion recognition relies on autonomous or involuntary physiological responses, including electrocardiography (ECG), galvanic skin response (GSR), and central nervous system signals such as EEG and fNIRS [4–6]. These signals can partially bypass subjective concealment and more objectively reflect changes in emotional arousal and valence, making them particularly suitable for reliability-critical applications such as mental-health assessment and brain-computer interfaces [7].

Early studies on physiological-signal-based emotion recognition predominantly focused on a single modality. Among various physiological signals, EEG provides millisecond-level temporal resolution and has been widely used to characterize transient neural oscillatory activity under affect elicitation. Complementarily, fNIRS measures concentration changes in oxy-hemoglobin (HbO) and deoxy-hemoglobin (HbR) to indirectly reflect local cortical blood flow and metabolic activity; although its temporal resolution is lower, it offers superior spatial localization. Under the mechanism of neurovascular coupling, EEG and fNIRS provide two complementary observation views of the same underlying neural activity, and both can be acquired using portable wearable devices [8]. Therefore, compared with either modality alone, EEG–fNIRS multi-modal fusion is expected to achieve substantial gains in accuracy and robustness in emotion recognition.

In recent years, graph neural networks (GNNs) have emerged as a class of deep learning methods for non-Euclidean data, learning representations on graph-structured inputs primarily via message passing and neighborhood aggregation, and have been increasingly applied to emotion recognition and related tasks [9]. EEG and fNIRS can be naturally represented as graphs by treating measurement channels as nodes and functional connectivity as edges, enabling the mining of emotion-related network features at the node, edge, and graph levels, which also improves interpretability [10]. Jia et al. [11] proposed a channel-relation graph convolutional network (CR-GCN) and achieved classification accuracy rates of 94.69% and 93.95% for valence and arousal on the DEAP dataset using EEG. Li et al. [12] introduced an attention-based temporal graph representation network (ATGRNet) and obtained 92.59% accuracy on the SEED dataset with EEG signals. For EEG–fNIRS fusion-based emotion recognition, Chen et al. [13] developed a model combining GCNs with a capsule-attention network (GCN-CA-CapsNet) and reported 97.91% accuracy, while Hou et al. [14] constructed causal-coupled brain networks in source space and employed SVM classification, achieving 96.6% accuracy.

Despite the impressive performance reported under controlled laboratory settings, cross-subject emotion recognition for practical deployment still faces substantial challenges. First, inter-individual variability is a primary bottleneck for generalization. Due to inherent differences in anatomical structures, baseline physiological states, and affective response patterns, models trained with conventional cross-entropy losses often exhibit significant performance degradation on unseen subjects. Many high-accuracy approaches rely on subject-dependent training paradigms, with which it is difficult to satisfy the plug-and-play requirement of real-world BCIs. For example, Chang et al. [15] proposed a multi-scale hyperbolic contrastive learning network (MSHCL) and achieved 89.3% cross-subject accuracy on the SEED dataset. Si et al. [16] designed a dual-branch joint network (DBJNet) and reached 74.8% cross-subject three-class accuracy on a self-built fNIRS database. These studies suggest that cross-subject performance remains markedly lower than subject-dependent performance, largely because EEG feature distributions exhibit substantial domain shift across subjects, making it difficult for classifiers to learn universal affective representations [17]. Second, semantic alignment between heterogeneous modalities is still insufficient.

Although EEG and fNIRS are physiologically related, their feature distributions lie on fundamentally different manifolds. Existing multi-modal fusion methods are often limited to data-level concatenation or shallow feature-level fusion; such straightforward combinations typically fail to eliminate cross-modal heterogeneity, leading to inefficient exploitation of complementary information [18]. Third, conventional GNNs often adopt isotropic aggregation when integrating neighborhood information and lack attention mechanisms that dynamically adjust inter-regional connection weights for emotion-related tasks, thereby limiting the ability to capture key affective brain networks [19]. Inspired by recent advances in computer vision and graph representation learning [20–23], we introduce contrastive learning into EEG–fNIRS fusion-based emotion recognition to better represent and relate the two signals.

Accordingly, we propose a dual-contrastive cross-modal attentional graph isomorphism network (DC-AGIN), which aims to achieve deep disentanglement and alignment of representations via two complementary contrastive learning mechanisms: cross-modal contrastive learning and cross-subject supervised contrastive learning. Architecturally, an attention graph isomorphism network (AGIN) is employed as the encoder to enhance the model's capability to focus on critical brain network topological connections. From the learning-strategy perspective, cross-modal contrastive learning maps heterogeneous EEG and fNIRS signals into a shared semantic subspace to maximize mutual information across modalities; meanwhile, cross-subject supervised contrastive learning explicitly suppresses subject-specific identity information and encourages the encoder to concentrate on intrinsic emotion-class characteristics, thereby substantially improving generalization to unseen subjects. Our main contributions are summarized as follows:

1. We integrate a graph attention mechanism into the graph isomorphism network to improve neighborhood aggregation by dynamically assigning weights according to the importance of each neighboring node.
2. We leverage cross-modal contrastive learning to align the semantic information between EEG and fNIRS, thereby mitigating cross-modal heterogeneity. In addition, cross-subject contrastive learning is employed on the fused EEG–fNIRS representations to facilitate cross-subject emotion recognition and reduce inter-subject discrepancies in affective patterns.
3. Extensive experiments on our self-collected EEG and fNIRS dataset demonstrate that the proposed method achieves state-of-the-art performance.

2. Data and Pre-Built Model

2.1. Data Acquisition and Preprocessing

2.1.1. Subjects and Data Collection

A total of 30 healthy participants (15 males and 15 females) were recruited, with a mean age of 24.1 ± 1.7 years. All participants had normal (or corrected-to-normal) vision and normal hearing, were right-handed, and had no history of neurological or psychiatric disorders. Before each experiment, participants were fully informed of the study objectives and procedures. Written informed consent was obtained from all participants, and each participant received monetary compensation after completing the experiment. The study protocol complied with the Declaration of Helsinki and was approved by the Research Ethics Committee of Taiyuan University of Technology. For data acquisition, whole-brain EEG signals were recorded using an ESI-NeuroScan system with 62 channels at a sampling rate of 1000 Hz. Simultaneously, fNIRS signals were synchronously collected using a portable NirsSmart fNIRS system at a sampling rate of 11 Hz. The optode layout covered the frontal and temporal regions, comprising 18 measurement channels.

2.1.2. Experimental Paradigm

All experimental data were collected in an electromagnetically shielded room, and the acquisition workflow is illustrated in Figure 1a. During the experiment, participants sat comfortably in front of a monitor and were instructed to remain as still as possible, minimizing body movements and eye blinks to reduce artifact contamination. A video-elicitation paradigm was adopted. In total, 60 video clips (1–2 min each) were selected as affective stimuli, covering four emotional states: sadness, happiness, neutral, and fear (15 clips per category). The video clips were pre-evaluated in terms of arousal and valence based on the Self-Assessment Manikin (SAM) scores provided by 20 individuals who did not participate in the main experiment, and the final affective video stimuli were selected according to these ratings. After each clip, participants were required to complete a self-assessment questionnaire within 30 s to verify whether the target emotion was successfully elicited. Each participant completed 60 experimental trials. After the experiment, the system recorded the corresponding 62-channel EEG data and 18-channel fNIRS data; the spatial distribution of measurement channels is shown in Figure 1b.

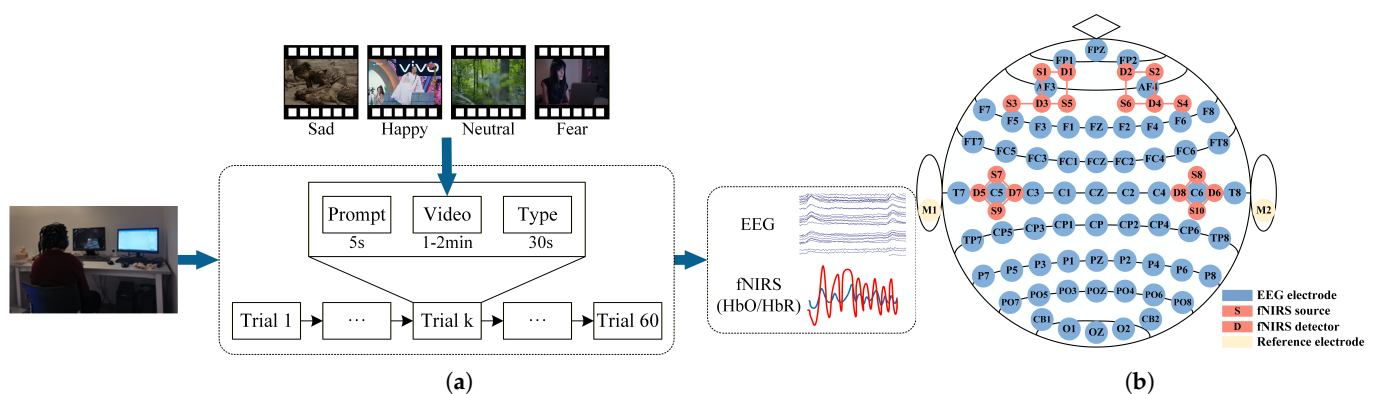


Figure 1. Experimental procedure and EEG–fNIRS channel configuration. (a) Data acquisition process; (b) EEG and fNIRS channel distribution diagram.

2.1.3. Data Preprocessing

Given that raw EEG signals are susceptible to contamination from electrooculography (EOG), electromyography (EMG), and various noise sources, EEG preprocessing was performed using the EEGLAB v2023.0 toolbox. Specifically, the bilateral mastoids (M1 and M2) were used as re-referencing electrodes. A Hamming-windowed sinc FIR filter was applied for band-pass filtering in the range of 0.5–45 Hz. To remove signal drift, baseline correction was performed by subtracting the mean value within the 2 s pre-stimulus interval. Independent component analysis (ICA) was then employed to remove blink/ocular and myogenic artifact components; blink/ocular components typically exhibit prominent frontal weights on the scalp topography or abrupt large-amplitude fluctuations in the time course, whereas EMG artifacts are often characterized by elevated high-frequency energy in the waveform. These operations were used solely to denoise the signals and did not involve discarding any data segments. The EEG data were subsequently downsampled to 200 Hz and segmented into 1 s samples. For fNIRS signals, a 0.01–0.2 Hz band-pass filter was applied to suppress physiological drift and high-frequency noise, and the same 2 s baseline correction was performed. For fNIRS signals, a 2 s baseline correction was first conducted, and a 0.01–0.2 Hz band-pass filter was applied to suppress physiological drift and high-frequency noise. Subsequently, hemodynamic analysis was performed using NirSpark v2.1 software, where changes in optical density were converted into relative concentration changes in HbO and HbR according to the modified Beer–Lambert law [24].

The preprocessed fNIRS data retained the original sampling rate of 11 Hz and were likewise segmented into 1 s epochs aligned with the EEG epochs.

2.1.4. Feature Extraction and Graph Construction

After EEG preprocessing, the signals were decomposed into five canonical frequency bands: δ (0.5–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz), and γ (30–45 Hz). For each band, the differential entropy (DE) feature was computed [25], which can be expressed as

$$DE = - \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) dx = \frac{1}{2} \log(2\pi e \sigma^2), \quad (1)$$

here x denotes the EEG signal amplitude of a given channel within a specific frequency band and time window, which is assumed to follow a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Therefore, for each EEG channel, a 5-dimensional feature vector was obtained, corresponding to the energy-related DE features of the five frequency bands.

The fNIRS signals primarily reflect the temporal evolution of cerebral hemodynamics. To capture the dynamic characteristics of HbO and HbR concentrations during affect elicitation and to keep the feature dimensionality consistent with the EEG representation for subsequent unified node processing, we extracted five statistical features (mean, variance, skewness, maximum, and minimum) from the HbO and HbR time series of each fNIRS channel. These features characterize both the statistical properties and variation trends of fNIRS signals and are defined as

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y(i), \sigma^2(y) = \frac{1}{N} \sum_{i=1}^N (y(i) - \bar{y})^2, s(y) = \frac{1}{N\sigma^3} \sum_{i=1}^N (y(i) - \bar{y})^3, \quad (2)$$

$$MIN = \min_{i=1, \dots, N} y(i), MAX = \max_{i=1, \dots, N} y(i).$$

Here, $y(i)$ denotes the i -th fNIRS sample, and N is the sample length; the skewness s reflects the asymmetry of the data distribution. Finally, each EEG electrode and each fNIRS channel are represented as a feature vector in $\mathbb{R}^5$.

2.1.5. Graph Construction

For EEG signals, we define the 62 EEG electrodes as a node set V_{EEG} with $|V_{\text{EEG}}| = 62$. We compute the correlations among all EEG channel signals using the Pearson correlation coefficient (PCC) [26] in a subject-wise manner to avoid any training–test leakage. Specifically, for each subject, the PCC is estimated using only the trials from that subject, yielding a subject-specific adjacency matrix $\mathbf{A}_{\text{EEG}}$. The PCC measures the linear association between two EEG channel signals, i.e., it quantifies how consistently the amplitudes of two channel-wise feature vectors co-vary. The magnitude $|\text{PCC}|$ indicates the strength of the correlation: values closer to 1 correspond to stronger linear dependence, whereas $|\text{PCC}| = 0$ indicates that the two vectors are linearly uncorrelated. According to research by Koszut et al. [27], we then define connections with $|\text{PCC}| > 0.5$ as edges $\mathcal{E}_{\text{EEG}}$ to retain only sufficiently correlated channel pairs and suppress weak/noisy couplings. In this way, an undirected EEG graph is constructed as $\mathcal{G}_{\text{EEG}} = (V_{\text{EEG}}, \mathcal{E}_{\text{EEG}})$, together with the node-feature matrix $\mathbf{X}_{\text{EEG}} \in \mathbb{R}^{4845 \times 62 \times 5}$ and the subject-specific adjacency matrix $\mathbf{A}_{\text{EEG}} \in \mathbb{R}^{62 \times 62}$, where each node feature corresponds to the DE feature computed above. Here, 4845 denotes the total number of samples obtained after segmenting the signals into 1 s windows.

$$\text{PCC}_{x,y} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\mathbb{E}[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y}. \quad (3)$$

Here, $\text{cov}(x, y)$ denotes the covariance, and σ_x, σ_y are the standard deviations of the two channel signals. According to Achard et al. [26], we retain, for each node, the top 20% strongest connections when building the graph.

For fNIRS signals, we perform graph-level fusion of HbO and HbR. Specifically, the 18-channel HbO signals and the 18-channel HbR signals are jointly treated as graph nodes, yielding a node set V_{fNIRS} with $|V_{\text{fNIRS}}| = 36$. Following the same procedure as for EEG, we compute PCC-based correlations to construct an fNIRS feature matrix $\mathbf{X}_{\text{fNIRS}} \in \mathbb{R}^{4845 \times 36 \times 5}$ and an adjacency matrix $\mathbf{A}_{\text{fNIRS}} \in \mathbb{R}^{36 \times 36}$.

Figure 2 presents the adjacency matrices of the constructed graphs for the two modalities. In summary, constructing modality-specific graphs enables subsequent GNNs to mine comprehensive cortical electrophysiological and neurovascular characteristics within a unified topological space.

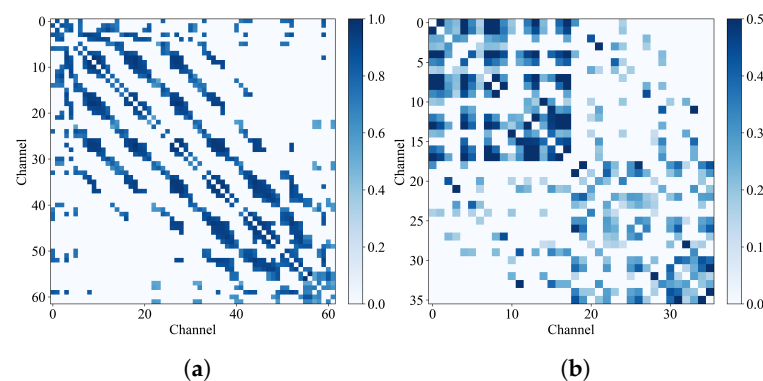


Figure 2. EEG and fNIRS graph adjacency matrices: (a) EEG adjacency matrix; (b) fNIRS adjacency matrix.

3. Methods

3.1. Graph Isomorphism Network Based on Graph Attention

Among many GNN variants, the graph isomorphism network (GIN) has been theoretically shown to possess strong graph-discriminative capability [23]. Under mild assumptions (e.g., when node and edge attributes are not overly complex), the GIN can achieve expressiveness comparable to the Weisfeiler–Lehman (WL) graph isomorphism test [28], thereby effectively distinguishing different non-isomorphic graph structures. However, the standard GIN adopts an isotropic neighbor-aggregation scheme, implicitly assuming that all neighbors of a central node contribute equally.

In EEG–fNIRS emotion recognition, functional brain connectivity exhibits pronounced dynamics and heterogeneity: specific affective states tend to activate particular neural circuits rather than uniformly engaging the whole brain. Motivated by this observation and inspired by GAT [29], we propose an improved Attention-GIN (AGIN) module as the encoder. While preserving the strong topological discriminability of the GIN, the AGIN introduces an attention mechanism to enable anisotropic feature aggregation, adaptively focusing on informative neighboring nodes and aggregating salient features more effectively.

Given a brain network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, let $\mathbf{h}_i^{(l)}$ denote the feature vector of node v_i at the l -th layer. The AGIN consists of two steps: attention-coefficient computation and weighted message passing. Specifically, the attention score between the central node v_i and its neighbor $v_j \in \mathcal{N}(i)$ is first computed as

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}\mathbf{h}_i^{(l)} \parallel \mathbf{W}\mathbf{h}_j^{(l)} \right]\right), \quad (4)$$

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})}. \quad (5)$$

To model interactions among nodes, we concatenate the transformed features of the central node and its neighbor and then compute the raw attention score e_{ij} via a learnable linear transformation $\mathbf{W}$ and LeakyReLU activation, where $\mathbf{a}$ is a learnable attention weight vector. Subsequently, a Softmax function is applied to normalize the attention scores over all neighbors, yielding the attention coefficient α_{ij} , which adaptively reflects the importance of neighbor v_j to the central node v_i . After obtaining the attention weights, we further modify the aggregation rule of the GIN. The standard GIN aggregation can be written as

$$\mathbf{h}_i^{(l+1)} = \text{MLP}^{(l)} \left((1 + \epsilon^{(l)}) \mathbf{h}_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \mathbf{h}_j^{(l)} \right). \quad (6)$$

Here, ϵ is a learnable parameter (or a fixed constant) that controls the contribution of the central node's own feature, and MLP denotes a multi-layer perceptron used to model nonlinear transformations. When aggregating neighborhood information, we weight each neighbor by its attention coefficient α_{ij} , as illustrated in Figure 3, which shows the attention computation and the overall AGIN framework. After obtaining the attention coefficients between a central node and its neighbors, the next-layer node representation is computed via weighted summation followed by an MLP. The AGIN aggregation is defined as

$$\mathbf{h}_i^{(l+1)} = \text{MLP}^{(l)} \left((1 + \epsilon^{(l)}) \mathbf{h}_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{h}_j^{(l)} \right). \quad (7)$$

With this design, the AGIN can adaptively emphasize informative connections and suppress noisy links according to physiological-signal characteristics under affect elicitation, thereby strengthening information flow over key brain regions and producing more discriminative graph representations. Notably, the standard GAT performs attention-driven neighborhood aggregation, whereas the AGIN superimposes attention reweighting on top of the GIN-style MLP and sum aggregation, thereby preserving graph isomorphism expressive power. We adopt a two-layer AGIN encoder: the first layer maps the input from 5 to 16 dimensions, and the second layer maps the representation from 16 to 32 dimensions. The graph-level features obtained by aggregating node embeddings from these two layers are used for classification and subsequent processing. In addition, to alleviate the problem of vanishing or exploding gradients, we introduce residual connections. $\mathbf{H}$ denotes the graph features output by the encoder.

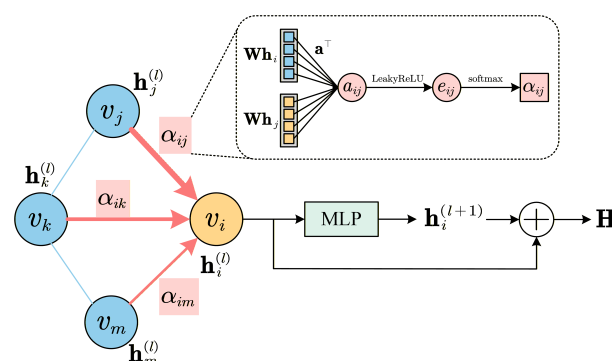


Figure 3. Attention mechanism and AGIN convolution.

3.2. EEG-fNIRS Cross-Modal Contrastive Learning

Although EEG and fNIRS signals are recorded simultaneously to capture neural activity during the same affect elicitation process, they exhibit pronounced heterogeneity: EEG reflects millisecond-level high-frequency electrophysiological fluctuations, whereas fNIRS manifests second-level low-frequency hemodynamic variations. Such substantial differences in physical attributes lead to feature distributions that reside on largely non-overlapping manifolds, making it difficult for simple concatenation to effectively exploit cross-modal complementarity.

Inspired by the success of CLIP [22] in vision–language representation learning, we transfer the contrastive learning paradigm to multi-modal brain-signal analysis. Similar to CLIP, which aligns image and text features via a dual-encoder framework, we design an EEG–fNIRS cross-modal contrastive learning module. Specifically, we employ the proposed AGIN as modality-specific encoders for EEG and fNIRS, respectively, and use contrastive learning to align the two heterogeneous physiological signals by projecting them into a unified latent semantic space.

First, the two constructed graphs $\mathcal{G}_{\text{EEG}}$ and $\mathcal{G}_{\text{fNIRS}}$ are fed into two non-shared encoders, $f_{\text{EEG}}(\cdot)$ and $f_{\text{fNIRS}}(\cdot)$, each implemented by a stacked AGIN. For the i -th sample in a mini-batch containing N samples, the resulting high-dimensional graph-level representations are

$$\mathbf{H}_{\text{EEG}}^{(i)} = f_{\text{EEG}}(\mathcal{G}_{\text{EEG}}^{(i)}), \quad \mathbf{H}_{\text{fNIRS}}^{(i)} = f_{\text{fNIRS}}(\mathcal{G}_{\text{fNIRS}}^{(i)}), \quad (8)$$

where $\mathbf{H} \in \mathbb{R}^{d_h}$ and d_h denotes the dimensionality of the encoder output.

Prior studies [20] have shown that directly applying contrastive loss on the encoder outputs may discard information that is beneficial for downstream classification. Therefore, we introduce an independent nonlinear projection head $p(\cdot)$ after each encoder to map the high-dimensional representation $\mathbf{H}$ onto a lower-dimensional hypersphere for similarity computation. The projection head is designed as a single-hidden-layer multi-layer perceptron (MLP), consisting of a fully connected layer, a ReLU activation, and a linear layer, followed by ℓ_2 normalization. Taking the EEG branch as an example, the projected feature $\mathbf{z}_{\text{EEG}}^{(i)}$ is computed as

$$\mathbf{z}_{\text{EEG}}^{(i)} = p(\mathbf{H}_{\text{EEG}}^{(i)}) = \text{Normalize}(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{h}_{\text{EEG}}^{(i)} + \mathbf{b}_1) + \mathbf{b}_2), \quad (9)$$

where $\mathbf{W}$ and $\mathbf{b}$ denote the weight matrices and bias vectors of the MLP. The input and output dimensions of $\mathbf{W}_1$ are 32 and 32, respectively, while those of $\mathbf{W}_2$ are 32 and 128; together, they constitute a two-layer fully connected network. The fNIRS branch is computed in the same manner to obtain $\mathbf{z}_{\text{fNIRS}}^{(i)}$. The resulting $\mathbf{z}$ vectors are then used to compute the contrastive loss and the corresponding pull–push distances.

To achieve feature alignment, we introduce a symmetric cross-modal contrastive loss in the feature-fusion stage, following the InfoNCE [30] formulation. The loss consists of two parts: (i) using EEG as the anchor to retrieve the corresponding fNIRS sample, denoted by $\mathcal{L}_{E \rightarrow f}$, and (ii) using fNIRS as the anchor to retrieve the corresponding EEG sample, denoted by $\mathcal{L}_{f \rightarrow E}$. For a mini-batch of size N , let $\mathbf{z}_E^{(i)}$ and $\mathbf{z}_f^{(i)}$ be the ℓ_2 -normalized projected features of the i -th sample from EEG and fNIRS, respectively. We define positive pairs as matched EEG–fNIRS views from the same sample, i.e., $\{(\mathbf{z}_E^{(i)}, \mathbf{z}_f^{(i)})\}_{i=1}^N$, while the remaining $N(N-1)$ mismatched pairs $\{(\mathbf{z}_E^{(i)}, \mathbf{z}_f^{(j)})\}_{i \neq j}$ are treated as negatives. With this construction, the model is encouraged to identify correct cross-modal correspondences across all samples,

thereby capturing shared affect elicitation patterns between the two modalities. The loss is defined as follows:

$$\mathcal{L}_{E \rightarrow f} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{z}_E^{(i)}, \mathbf{z}_f^{(i)}) / \tau_m)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{z}_E^{(i)}, \mathbf{z}_f^{(j)}) / \tau_m)}. \quad (10)$$

$$\mathcal{L}_{f \rightarrow E} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{z}_f^{(i)}, \mathbf{z}_E^{(i)}) / \tau_m)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{z}_f^{(i)}, \mathbf{z}_E^{(j)}) / \tau_m)}. \quad (11)$$

$$\mathcal{L}_{CM} = \frac{1}{2} (\mathcal{L}_{E \rightarrow f} + \mathcal{L}_{f \rightarrow E}). \quad (12)$$

Here, $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$ denotes the cosine similarity, and τ_m is the temperature parameter (set to 0.1 in this paper) used to adjust the model's discrimination sharpness over difficult samples. Specifically, $\mathcal{L}_{E \rightarrow f}$ maximizes the cosine similarity between each EEG sample and its corresponding fNIRS sample; symmetrically, $\mathcal{L}_{f \rightarrow E}$ maximizes the cosine similarity between each fNIRS sample and its corresponding EEG sample. The final cross-modal contrastive loss is the average of these two directional losses, and the overall cross-modal contrastive learning pipeline is illustrated in Figure 4.

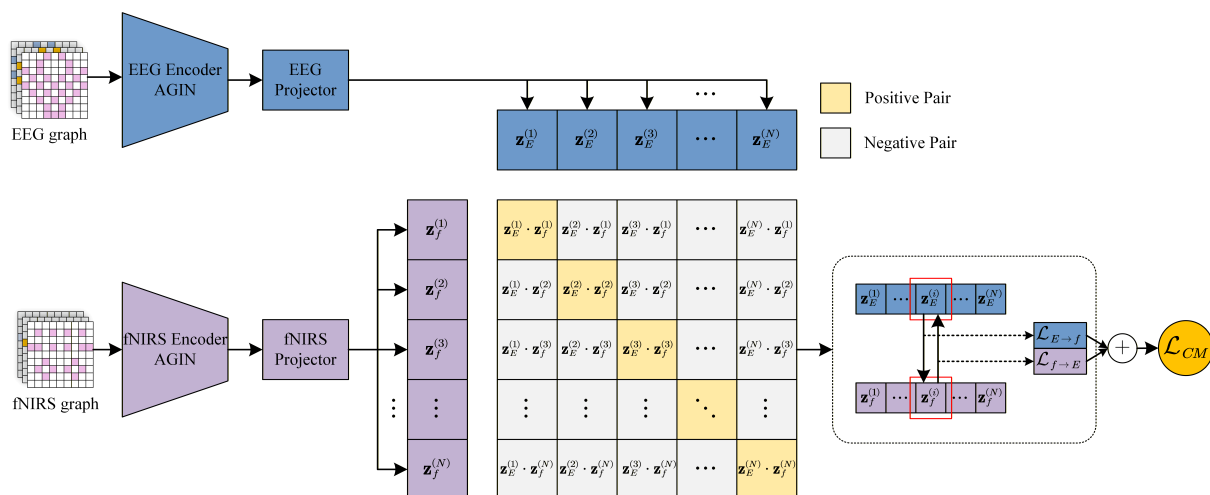


Figure 4. EEG–fNIRS cross-modal contrastive learning pipeline.

In Figure 4, the red boxes represent the two paired modal samples. By jointly optimizing the two directions and minimizing $\mathcal{L}_{CM}$, the model ensures that regardless of which modality is used as the query, the corresponding feature in the other modality can be accurately retrieved within the shared semantic space. This yields more robust semantic alignment and enables the model to learn affect-relevant structures shared across modalities (e.g., consistent affect elicitation patterns), thereby alleviating the distribution mismatch induced by heterogeneous data.

After alignment, the two modality-specific vectors become close to each other in the shared semantic space and tend to point in similar directions (indicating consistent affective semantics). We then fuse them by simple averaging to obtain the cross-modal representation $\mathbf{z}_F$, which is subsequently ℓ_2 -normalized and fed into downstream classification. The fusion is defined as

$$\mathbf{z}_F = \text{Normalize}\left(\frac{\mathbf{z}_E + \mathbf{z}_f}{2}\right). \quad (13)$$

3.3. Cross-Subject Supervised Contrastive Learning

Traditional supervised learning methods (e.g., cross-entropy loss $\mathcal{L}_{CE}$) mainly focus on learning decision boundaries that separate emotion classes on the training set. However, in the leave-one-subject-out (LOSO) setting, such approaches often cause the model to overfit subject-specific physiological patterns rather than learning truly subject-invariant affective characteristics. When encountering an unseen test subject, these idiosyncratic patterns no longer hold, resulting in a substantial performance drop.

To address this generalization bottleneck, we propose a cross-subject supervised contrastive learning strategy. The key idea is to explicitly perform “feature disentanglement” in the representation space: suppress subject-identity information while enhancing emotion-discriminative information.

In standard supervised contrastive learning (SupCon) [21], samples sharing the same emotion label are treated as positive pairs. This is insufficient for cross-subject tasks because it may simply pull together samples from the same subject. Therefore, we define a cross-subject positive-pair construction. Consider a training mini-batch containing N fused samples. For an anchor sample i , with representation, emotion label, and subject identity denoted by $\mathbf{z}_i$, y_i , and s_i , respectively, the positive set $\mathcal{P}(i)$ is defined by requiring the same emotion label ($y_j = y_i$) but a different subject identity ($s_j \neq s_i$):

$$\mathcal{P}(i) = \{j \in \{1, \dots, N\} \mid j \neq i, y_j = y_i, s_j \neq s_i\}. \quad (14)$$

Two samples are then drawn from $\mathcal{P}(i)$ to form the positive set for anchor i . Correspondingly, all samples in the mini-batch that do not satisfy the above conditions (i.e., samples with different emotion labels) are paired with the anchor to construct the negative set. It is worth noting that we do not use the sample itself to form a positive pair. This is because self-pairs do not provide additional cross-subject information for the same emotion class; excluding them explicitly blocks the shortcut of leveraging subject-identity cues to reduce contrastive distances.

Based on the above positive/negative sample construction, the cross-subject supervised contrastive loss is defined as

$$\mathcal{L}_{CS} = \sum_{i=1}^N \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau_s)}{\sum_{a \in \mathcal{A}(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau_s)}. \quad (15)$$

Here, $\mathcal{A}(i) = \{j \mid j \neq i\}$ denotes the set of all samples in the batch except anchor i , and $|\mathcal{P}(i)|$ is the number of positive samples. The parameter τ_s is the cross-subject temperature coefficient (set to 0.05 in this paper). By minimizing $\mathcal{L}_{CS}$, the model pulls together samples from different subjects but with the same emotion label, encouraging their feature distributions to overlap while pushing apart samples from different emotion classes. This yields representations that are more robust to inter-subject variability and substantially improves generalization to unseen subjects. Algorithm 1 provides the pseudocode for computing the cross-subject contrastive loss.

Algorithm 1 Cross-Subject supervised contrastive learning

Input: A mini-batch of fused features $\{\mathbf{z}_i\}_{i=1}^N$; emotion labels $\{y_i\}$; subject IDs $\{s_i\}$; temperature τ_s

Output: Cross-subject supervised contrastive loss $\mathcal{L}_{CS}$ (Equation (15))

- 1: Normalize features: $\mathbf{z}_i \leftarrow \mathbf{z}_i / \|\mathbf{z}_i\|_2, \forall i$
- 2: Compute similarity matrix: $S_{ij} \leftarrow \mathbf{z}_i^\top \mathbf{z}_j / \tau_s, \forall i, j$
- 3: Mask self-pairs: set $S_{ii} \leftarrow -\infty, \forall i$
- 4: $L \leftarrow 0, C \leftarrow 0$
- 5: **for** $i \leftarrow 1$ **to** N **do**
- 6: Construct cross-subject positive set $P(i)$ according to Equation (14)
- 7: Construct all-others set $A(i) \leftarrow \{j \mid j \neq i\}$
- 8: **if** $|P(i)| > 0$ **then**
- 9: Compute anchor loss L_i according to Equation (15) using $P(i)$ and $A(i)$
- 10: $L \leftarrow L + L_i$
- 11: $C \leftarrow C + 1$
- 12: **end if**
- 13: **end for**
- 14: $\mathcal{L}_{CS} \leftarrow L / \max(C, 1)$
- 15: **return** $\mathcal{L}_{CS}$

Note that C counts the number of valid anchors that admit at least one cross-subject positive. In rare cases, C can be 0 when a mini-batch contains no emotion class represented by at least two different subjects (e.g., all samples of a class come from a single subject, or some classes appear only once in the batch). We, therefore, use $\max(C, 1)$ to avoid division by zero; when $C = 0$, the cross-subject contrastive loss is effectively set to zero for that batch, and only the remaining objectives contribute to the update.

3.4. Overall Model

To translate the extracted high-level semantic features into the final emotion predictions, we attach a linear classifier after the feature-fusion module. Specifically, the input is the ℓ_2 -normalized fused multi-modal feature vector, and the classifier is implemented by a fully connected layer. The cross-entropy classification loss is defined as

$$\mathcal{L}_{CLS} = -\frac{1}{N} \sum_{n=1}^N \sum_{i=1}^C l'_{n,i} \log(p_{n,i}), \quad (16)$$

where N is the batch size, C is the number of emotion classes, $l'_{n,i} \in \{0, 1, 2, 3\}$ denotes the one-hot ground-truth label of the n -th sample with $\sum_{i=1}^C l'_{n,i} = 1$, and $p_{n,i}$ is the predicted probability for class i .

As illustrated in Figure 5, the overall objective combines the above losses into a multi-task joint optimization target. The total loss is formulated as a weighted sum of three terms:

$$\mathcal{L}_{total} = \mathcal{L}_{CLS} + \lambda_1 \mathcal{L}_{CM} + \lambda_2 \mathcal{L}_{CS}. \quad (17)$$

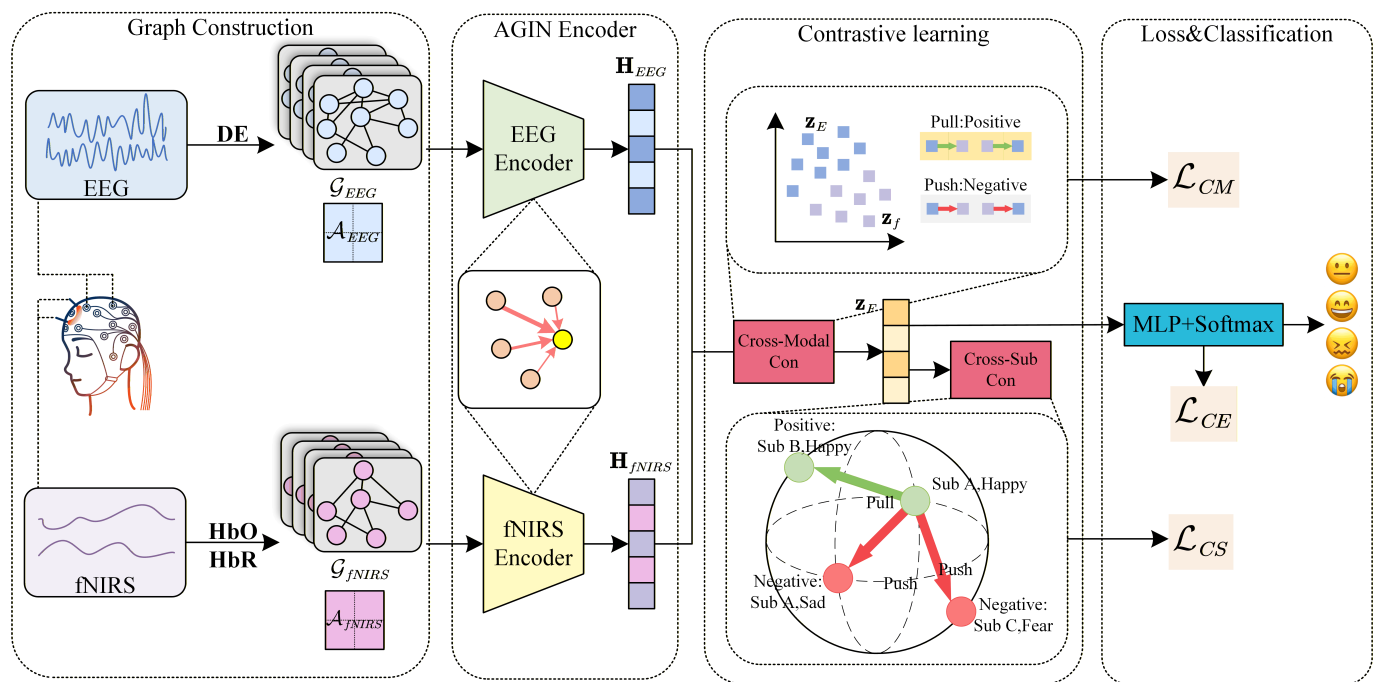


Figure 5. DC-AGIN-based EEG–fNIRS cross-subject emotion recognition: Modality-specific graph representations of EEG and fNIRS are extracted using AGIN encoders. Cross-modal contrastive learning aligns heterogeneous modalities, while cross-subject contrastive learning improves generalization by suppressing subject-specific factors. The aligned features are fused and classified to produce final emotion predictions.

Here, $\mathcal{L}_{CLS}$ ensures discriminative representations for emotion classification, $\mathcal{L}_{CM}$ encourages consistent semantic representations between the EEG and fNIRS encoders to alleviate cross-modal heterogeneity, and $\mathcal{L}_{CS}$ suppresses subject-specific variations to learn subject-invariant affective representations, thereby improving cross-subject generalization. The hyperparameters λ_1 and λ_2 balance the contribution of the contrastive objectives against the primary classification objective. By minimizing $\mathcal{L}_{total}$, the model can be trained end-to-end to jointly optimize representation learning and classification performance.

4. Experimental Results and Analysis

4.1. Experimental Setup

To evaluate the effectiveness of the proposed method, experiments were conducted on our self-collected EEG–fNIRS emotion dataset TYUT3.0. Data from 30 participants across four emotions (sadness, happiness, neutral, and fear) were used for training, validation, and testing. Existing multi-modal EEG-based emotion recognition datasets [31] typically conduct validation with approximately 20–50 subjects; therefore, the sample size used in this work falls within the commonly adopted range in this field and is comparable to prior research. To rigorously assess emotion recognition performance under different experimental settings, we adopted two evaluation protocols:

1. **Subject-dependent setting:** For each participant, 5-fold cross-validation was performed by mixing and randomly shuffling samples from all subjects and then evenly splitting them into 5 folds. In each round, one fold was used as the test set. From the remaining four folds, one fold was further designated as the validation set, and the other three folds were used for training. The best hyperparameters were selected on the validation set, and the model checkpoint with the best validation performance was saved. The test set result was then taken as the performance for that round. Finally,

we report the mean accuracy (Acc) and standard deviation over the five rounds to evaluate the effectiveness of the encoder and cross-modal contrastive learning.

2. Subject-independent setting: We employed the leave-one-subject-out (LOSO) protocol. Since the dataset contains 30 subjects, the subject-independent experiments were conducted over 30 rounds. In each iteration, the data from one subject were held out as an independent test set, while the data from the remaining 29 subjects were used for model training. Within these 29 training subjects, we also performed a 5-fold cross-validation at the subject level to construct an inner validation split, where four folds were used for training and the remaining fold was used as the validation set for best-model selection. Z-score normalization was computed using statistics from the inner training split only, ensuring the independence of both the validation and test sets. The reported performance is the mean accuracy and standard deviation over the 30 test rounds, which is used to verify the effectiveness of cross-subject contrastive learning.

In addition, to assess the statistical significance of the performance improvements, we conducted paired-sample *t*-tests on the experimental results and reported the 95% confidence intervals (CIs). A *p*-value smaller than 0.05 was considered statistically significant. All experiments were conducted on an NVIDIA Tesla T4 GPU using the PyTorch 2.0 deep learning framework. The software environment included Python 3.11 and CUDA 11.8, and PyTorch Geometric was used to implement the GIN-based models. We optimized the model parameters using the cross-entropy loss and the Adam optimizer with a learning rate of 0.001. The batch size was 64, and the maximum number of training epochs was 100. An early-stopping strategy was adopted, whereby training was terminated if the validation accuracy did not improve for 20 consecutive epochs.

4.2. Results

4.2.1. EEG-fNIRS Fusion Emotion Recognition Comparison Experiment

To demonstrate the superiority of the proposed model, we compared it with several baseline methods, including the conventional machine-learning approach SVM [32] and representative GNN models such as GCN [33], GraphSAGE [34], GAT [29], and GIN [23]. We first considered the subject-dependent setting. In this case, we removed the final cross-subject contrastive learning term and kept only the AGIN encoder and the cross-modal contrastive learning module. The detailed five-fold cross-validation results are summarized in Table 1.

Table 1. Subject-dependent EEG-fNIRS fusion emotion recognition results.

Decoder	Sad	Happy	Neutral	Fear	Acc	95% CI
SVM [32]	89.1	91.0	91.8	89.54	90.36 ± 4.70	[88.60, 92.12]
GCN [33]	91.85	92.65	93.5	91.92	92.48 ± 4.12	[90.94, 94.02]
GraphSAGE [34]	92.7	93.44	94.06	92.8	93.25 ± 3.65	[91.89, 94.61]
GAT [29]	93.11	94.2	94.69	93.25	93.88 ± 3.40	[92.61, 95.15]
GIN [23]	93.95	94.84	95.34	94.15	94.57 ± 2.95	[93.47, 95.67]
DC-AGIN	96.36	97.26	97.76	96.54	96.98 ± 1.89	[96.27, 97.69]

As shown in Table 1, our method achieves a notably high recognition accuracy on the four-class task (Acc = 96.98%), outperforming the vanilla GIN baseline by approximately 2.41%. Based on the distribution of the collected results and the sample size, we additionally report the 95% CI of the accuracy for each method. Moreover, we conduct a paired *t*-test between DC-AGIN and the baseline GIN, yielding *p* < 0.01, which indicates that the proposed improvement is statistically significant under the subject-dependent setting. In

addition, among the compared GNN variants, the GIN provides the strongest baseline performance, supporting our choice of a GIN as the backbone for further improvements. These results indicate that when the distribution gap between training and test data is relatively small, the proposed DC-AGIN encoder can effectively capture fine-grained affective patterns from both EEG and fNIRS signals and align cross-modal representations, thereby improving subject-dependent emotion recognition performance.

To evaluate the model's adaptability to new users in practical brain-computer interface (BCI) applications, we re-introduced the cross-subject supervised contrastive learning module and conducted LOSO experiments using the complete DC-AGIN model. Under this setting, the test set consists entirely of data from a new subject who is not included in training, leading to a pronounced inter-subject discrepancy between the training and test sets. The subject-independent LOSO results are reported in Table 2.

Table 2. Subject-independent EEG-fNIRS fusion emotion recognition results.

Decoder	Sad	Happy	Neutral	Fear	Acc	95% CI
SVM [32]	45.23	50.11	49.37	47.61	48.08 ± 11.47	[43.80, 52.36]
GCN [33]	49.63	54.27	53.19	48.75	51.46 ± 11.10	[47.32, 55.60]
GraphSAGE [34]	51.97	53.83	55.29	48.19	52.82 ± 12.02	[48.33, 57.31]
GAT [29]	53.73	54.67	52.13	49.03	52.64 ± 11.98	[48.17, 57.11]
GIN [23]	51.47	55.03	54.27	51.71	53.12 ± 10.76	[49.10, 57.14]
DC-AGIN	62.93	63.87	64.99	59.45	62.56 ± 8.57	[59.36, 65.76]

As shown in Table 2, DC-AGIN maintains stable performance in the cross-subject setting. However, compared with subject-dependent five-fold cross-validation, the accuracy drops from 96.98% to 62.56%, indicating that cross-subject recognition is substantially more challenging than the subject-dependent setting. This gap mainly arises because, in the subject-dependent protocol, the training and test samples come from the same pool of subjects, allowing the model to exploit subject-specific physiological response patterns and thus achieve seemingly higher classification accuracy. In contrast, under the LOSO setting, the distribution shift of unseen subjects makes identity-related representations difficult to transfer directly, and the model may mistakenly treat inter-subject differences as emotion-related cues, leading to a pronounced degradation in generalization performance. Under LOSO evaluation, our model still substantially outperforms conventional methods and improves upon the vanilla GIN baseline by approximately 10.44%. This gain primarily stems from the proposed cross-subject supervised contrastive learning module, which effectively suppresses subject-specific identity noise and encourages the encoder to learn more generalizable affective patterns, thereby achieving smoother generalization to unseen subjects. Under the LOSO protocol with $n = 30$, we performed a paired t -test between DC-AGIN and the baseline GIN and obtained $p < 0.01$, which likewise indicates that the proposed improvement is statistically significant in the subject-independent setting.

In addition, we further compared our model with recent state-of-the-art techniques. Since most existing cross-subject emotion recognition studies focus on single-modality EEG and are evaluated on public datasets, to ensure a fair comparison with existing methods, we follow a minimal-adaptation principle when implementing the reproduced baselines. Specifically, we only unify the input interfaces (i.e., multi-modal channel numbers and feature dimensions), fusion strategy, and classification head so that each method can process the EEG-fNIRS multi-modal data in our study, without altering its core architecture or loss formulation. All baselines and DC-AGIN use exactly the same data splits and LOSO evaluation protocol, and hyperparameters are tuned within an identical search space. For each method, the final performance is obtained by selecting the best hyperparameters

based on the validation set within the training fold and then testing on the held-out subject. The comparative results are reported in Table 3.

Table 3. Subject-independent comparison results with existing methods.

Methods	Acc	Source of Results
GIN [23]	53.12 ± 10.76	Baseline Results
GCN-CapsNET [13]	53.91 ± 10.26	Original Results
MF-CNN [35]	58.59 ± 9.89	Original Results
DANN [36]	58.81 ± 10.51	Reproduced Results
CLISA [37]	60.78 ± 9.67	Reproduced Results
GC-STCL [38]	58.90 ± 10.25	Reproduced Results
CL-GNN [39]	59.97 ± 10.43	Reproduced Results
DC-AGIN	62.56 ± 8.57	Results of This Paper

Here, ref. [13] is an EEG–fNIRS framework based on a GCN and capsule networks, and [35] performs multi-modal emotion recognition by extracting causal features. However, most of these methods are designed around feature-network engineering and do not specifically optimize for cross-subject emotion recognition. In the LOSO evaluation, our DC-AGIN achieves an average accuracy of 62.56%, outperforming all compared methods. Although recent studies [36–38] have attempted to incorporate transfer or contrastive strategies and have achieved performance gains, DC-AGIN remains competitive under LOSO validation.

Compared with the domain-adversarial model DANN [36], which minimizes the distribution discrepancy between source- and target-domain data via adversarial learning, DC-AGIN improves accuracy by approximately 3.75%. Compared with the best-performing spatiotemporal representation learning model, CLISA [37], DC-AGIN achieves an additional gain of about 1.78%. Moreover, paired-sample *t*-tests indicate that all comparisons yield $p < 0.01$, confirming that the performance advantages of DC-AGIN are statistically significant.

These results further suggest that the proposed cross-modal contrastive learning is markedly more effective than other multi-modal fusion strategies. Although several competing methods are carefully designed to capture subject-specific characteristics, most of them are not explicitly optimized for cross-subject settings; consequently, under distribution shifts induced by unseen subjects, they tend to overfit the source-domain data and exhibit limited generalization. In addition, under the availability of explicit emotion labels and subject identities, the proposed cross-subject contrastive learning provides stronger cross-subject emotion recognition performance than domain adaptation approaches. Together with the alignment of EEG and fNIRS representations, this design plays a critical role in multi-modal cross-subject emotion recognition by enabling truly subject-invariant affective feature extraction, ultimately leading to improved recognition performance.

4.2.2. Ablation Experiment

To quantify the contribution of each key component in the DC-AGIN framework to cross-subject emotion recognition, we conducted a series of ablation studies under the LOSO protocol. The baseline model uses the vanilla GIN as the encoder, applies simple concatenation for multi-modal fusion, and is trained with the standard cross-entropy loss. AGIN replaces the GIN encoder with the attention-enhanced AGIN encoder only ($\lambda_1 = \lambda_2 = 0$). CM introduces only the cross-modal contrastive learning objective for feature alignment ($\lambda_1 = 0.1$, $\lambda_2 = 0$). CS introduces only the cross-subject supervised contrastive learning objective ($\lambda_1 = 0$, $\lambda_2 = 0.1$). DC-AGIN is the complete dual-contrastive

AGIN framework that jointly optimizes the three losses. Table 4 reports the LOSO results of all variants.

Table 4. Subject-independent model ablation results.

Model	Decoder	$\mathcal{L}_{CLS}$	$\mathcal{L}_{CM}$	$\mathcal{L}_{CS}$	Acc	95% CI
Baseline [23]	GIN	✓	✗	✗	53.12 ± 10.76	[49.10, 57.14]
DC-AGIN (AGIN)	AGIN	✓	✗	✗	54.26 ± 10.25	[50.43, 58.09]
DC-AGIN (CM)	GIN	✓	✓	✗	56.62 ± 9.94	[52.91, 60.33]
DC-AGIN (CS)	GIN	✓	✗	✓	60.04 ± 9.80	[56.38, 63.70]
DC-AGIN (Full)	AGIN	✓	✓	✓	62.56 ± 8.57	[59.36, 65.76]

Note: ✓ indicates the corresponding loss term is used; ✗ indicates it is not used.

Compared with the baseline (53.12%), the DC-AGIN (AGIN) variant, which introduces only the attention mechanism, improves the accuracy to 54.26%. This indicates that relative to the isotropic mean aggregation in the standard GIN, AGIN can adaptively assign higher weights to key brain regions for affective recognition, thereby filtering out task-irrelevant background noise to some extent and extracting more discriminative graph features.

By incorporating the cross-modal contrastive loss, DC-AGIN (CM) further increases the accuracy to 56.62%, yielding a 3.50% improvement over the baseline. This gain suggests that due to the heterogeneous physical characteristics of EEG and fNIRS signals, direct fusion alone is unlikely to be optimal. Cross-modal contrastive learning explicitly pulls temporally paired heterogeneous signals closer in the latent space, enabling semantic alignment and enhancing cross-modal complementarity.

Notably, introducing only the cross-subject supervised contrastive loss leads DC-AGIN (CS) to achieve an accuracy of 60.04%, corresponding to a 6.92% improvement—the largest gain among all single-module variants. In the ablation study, CS produces the largest gain, aligning with the key challenge in the subject-independent setting: strong inter-subject distribution shifts. Without explicitly suppressing identity-related cues, the model may overfit subject-specific patterns rather than emotion-relevant features; CS enforces subject-invariant representations by pairing same emotion, different subject positives, thus improving generalization the most. By contrast, CM mainly reduces the representation gap between EEG and fNIRS, improving cross-modal alignment and training stability, so its standalone LOSO gain is smaller but it strengthens robustness when combined with CS. Importantly, CM and CS are not additive: their effects partially overlap and the multi-loss constraints are coupled with diminishing returns, so the final improvement is smaller than summing individual gains.

Finally, the full DC-AGIN model, which integrates all components, achieves the best overall performance, reaching an average accuracy of 62.56% with a standard deviation reduced to 8.57 (the lowest among all variants). Moreover, paired *t*-tests between DC-AGIN and each ablated variant show that all comparisons yield $p < 0.01$, indicating that DC-AGIN significantly outperforms the ablated models. These results demonstrate that attention-based aggregation, cross-modal alignment, and cross-subject disentanglement are complementary rather than mutually exclusive. The complete framework not only improves recognition accuracy but also substantially enhances robustness when generalizing to unseen subjects. In addition, we recorded the computation time of the LOSO experiments for both the baseline and DC-AGIN. The total training time for each method was approximately 15 h, indicating a relatively high training cost, while the per-sample inference latency on the GPU was about 2.8 ms. We expect future work to further reduce the computational complexity; nonetheless, these results suggest that DC-AGIN improves emotion recognition performance without a substantial increase in computation time.

To evaluate the sensitivity of the proposed model to key hyperparameters, we further analyze the effects of the cross-modal contrastive temperature τ_m , the cross-subject supervised contrastive temperature τ_s , and the corresponding loss weights λ_1 and λ_2 on emotion recognition performance. Specifically, we conduct controlled experiments by varying τ over $\{0.05, 0.1, 0.5, 1\}$ and λ over $\{0.05, 0.1, 0.2, 0.5\}$. As shown in Figure 6, the performance varies moderately within a relatively wide parameter range, and the best results are obtained near the default configuration. In particular, when $\tau_m = 0.1$, $\tau_s = 0.05$, and $\lambda_1 = \lambda_2 = 0.1$, the model achieves the highest classification accuracy (92.65%), indicating that the proposed method exhibits reasonable robustness while benefiting from an appropriate choice of temperatures and loss-weight balancing. Further details are reported as follows:

1. **Temperature analysis:** The temperature τ_m is used in cross-modal contrastive learning, where positive and negative pairs are drawn from different modalities. Because cross-modal representations typically exhibit larger distribution shifts and more dispersed similarities, a relatively larger temperature (0.1) is adopted to smooth the logits, preventing an overly sharp similarity distribution from amplifying modality noise and causing unstable gradient updates. In contrast, τ_s is used in cross-subject supervised contrastive learning, where positive pairs satisfy the same emotion, different subject criterion. Within the shared fused representation space, the intra-class variations are comparatively controllable, yet stronger discriminative constraints are required to suppress subject-identity-related features. Therefore, a smaller temperature (0.05) is employed to strengthen the pull between same-class samples and the push between different-class samples.
2. **Loss-weight analysis:** The two weights λ_1 and λ_2 are introduced to balance the relative importance between the contrastive objectives and the primary classification objective. In other words, they control the trade-off among cross-modal semantic alignment, cross-subject-identity invariance, and emotion discriminability. When λ is too small, the contrastive losses are insufficient to mitigate modality discrepancies or inter-subject biases; when λ is too large, the contrastive objectives may dominate optimization and inadvertently weaken task-relevant discriminative cues for classification. Consequently, an appropriate weight (0.1) provides a better balance for cross-subject emotion recognition.

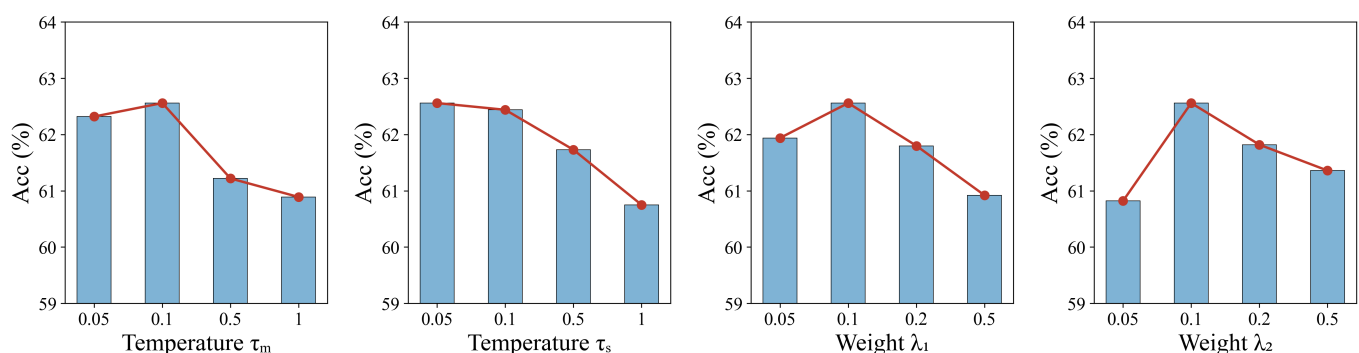


Figure 6. Effects of the contrastive learning temperature coefficient τ and the loss weight λ on the subject-independent emotion recognition performance of DC-AGIN.

To verify the effectiveness of EEG–fNIRS feature fusion and cross-modal contrastive learning, we conducted LOSO experiments using a single modality (EEG-only or fNIRS-only), and the results are reported in Table 5. The first two rows correspond to the uni-modal settings, where the model is trained and evaluated with only the AGIN encoder and the cross-subject contrastive learning module. As shown in the table, the multi-modal

modeling of EEG and fNIRS (including cross-modal contrastive learning and feature fusion) is necessary: compared with uni-modal EEG and uni-modal fNIRS, the multi-modal model improves the recognition accuracy by 2.97% and 20.11%, respectively. Moreover, paired-sample *t*-tests against the corresponding methods in the first two rows of Table 5 yield $p < 0.01$, indicating that the improvements are statistically significant. It can also be observed that the multi-modal gain over uni-modal fNIRS is substantially larger, mainly because HbO and HbR concentration variations evolve relatively slowly and thus provide limited discriminative power when used alone for affective representation. Overall, these modality ablation results demonstrate that the proposed model can effectively fuse the complementary affective cues from EEG and fNIRS, providing more comprehensive results.

Table 5. Subject-independent modality ablation results.

Modality	Sad	Happy	Neutral	Fear	Acc	95% CI
DC-AGIN (EEG)	58.90	61.20	62.45	57.75	60.03 ± 9.75	[56.39, 63.67]
DC-AGIN (fNIRS)	40.10	44.80	46.25	38.65	42.45 ± 17.22	[36.02, 48.88]
DC-AGIN (EEG+fNIRS)	62.93	63.87	64.99	59.45	62.56 ± 8.57	[59.36, 65.76]

4.2.3. Visualization Analysis

To visually verify the feature learning capability of DC-AGIN, we performed a comparative t-SNE visualization between the raw multi-modal features at the network input and the high-level semantic features obtained after the dual-stream encoders and the dual-contrastive learning scheme. The t-SNE hyperparameters were set as follows: the perplexity was set to 30, which lies within the commonly recommended range of 5–50 and provides a good trade-off between local and global structure; the random seed was fixed to 42 to ensure reproducibility; and the number of iterations was set to 1000 to facilitate convergence. As shown in Figure 7, the original feature space in (a) exhibits a highly entangled distribution: samples from different emotion classes overlap substantially and lack separable cluster structures. This reflects that the raw EEG–fNIRS signals contain considerable background noise and that the heterogeneous modalities suffer from a large distribution gap, resulting in poor linear separability. In contrast, the learned feature space in (b) shows a clear structural improvement. Previously scattered samples begin to form bounded manifold-like clusters according to affective semantics (e.g., Sad in blue and Happy in red become increasingly separable), and both intra-class compactness and inter-class separability are markedly enhanced. Quantitatively, the raw features yield a Silhouette score of 0.089 and a Davies–Bouldin index (DBI) of 3.847, indicating that the four emotion classes are highly entangled in the original feature space. In contrast, the fused features extracted by DC-AGIN achieve a higher Silhouette score of 0.312 and a lower DBI of 2.236, reflecting improved clustering quality and providing quantitative evidence that the proposed method effectively enhances separability among emotion classes. This transition from disordered to structured distributions strongly indicates that DC-AGIN can effectively suppress low-level physiological noise and map heterogeneous signals into a unified high-dimensional discriminative space, thereby providing a solid representation foundation for subsequent high-accuracy emotion classification.

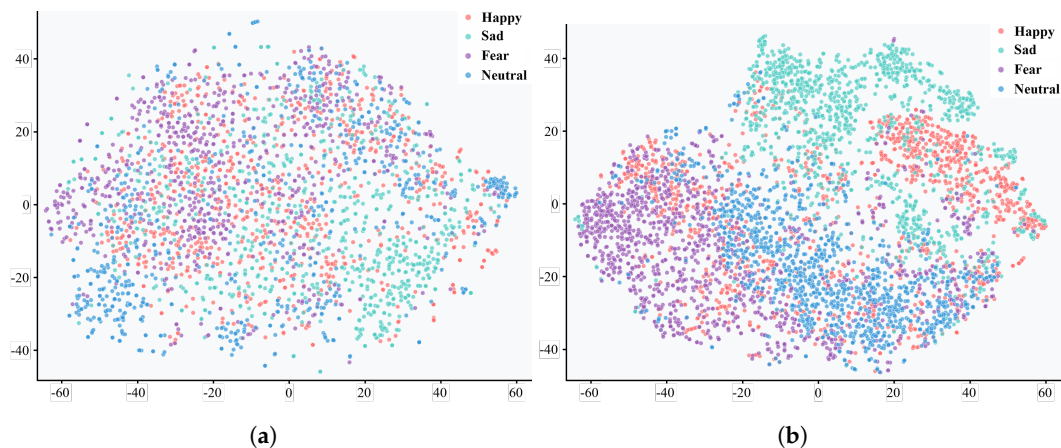


Figure 7. DC-AGIN model feature extraction t-SNE visualization: (a) original feature space; (b) after feature extraction by the network.

Figure 8 presents the confusion matrices of DC-AGIN under the subject-dependent (a) and subject-independent (b) evaluation protocols. Panel (a) indicates that the model achieves near-perfect classification performance, suggesting that when the training set contains partial data from the same participants, the model can readily capture subject-specific physiological patterns and learn clear decision boundaries. However, this idealized setting may mask the critical inter-subject variability challenge in affective computing. More practically relevant is the subject-independent LOSO performance shown in (b). Despite the distribution shift across unseen subjects, the model still preserves satisfactory discriminability, with an average diagonal accuracy of approximately 62.9%.

By examining the off-diagonal misclassification patterns, we find that the confusions are not random. A prominent mutual confusion occurs between Happy and Fear, which can be attributed to their high similarity in arousal; consequently, it is difficult for the model to distinguish these two high-energy affective states solely based on the intensity of physiological activation. Another major confusion arises between Sad and Neutral. Sadness is often associated with suppressed neural activity or low-energy states, which resemble the baseline Neutral condition in both EEG power spectra and hemodynamic responses. As a result, some sadness samples with weak fluctuations tend to be classified as the most stable neutral-like state.

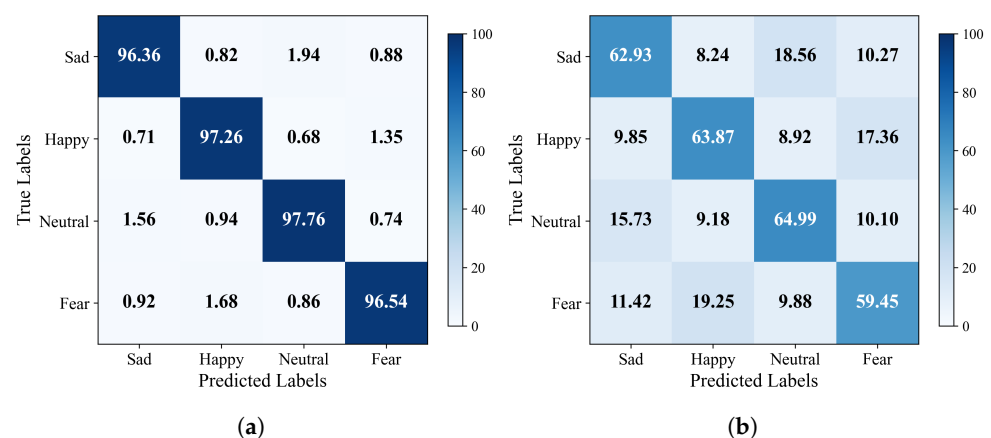


Figure 8. DC-AGIN model sentiment recognition confusion matrix: (a) 5-fold; (b) LOSO.

Figure 9 reports the per-subject emotion recognition performance under the LOSO protocol, where the blue and red curves correspond to the baseline method and the proposed

DC-AGIN, respectively. It can be observed that both methods exhibit noticeable fluctuations across different subjects, indicating that cross-subject emotion recognition is still affected by individual variability and differences in signal quality. Nevertheless, DC-AGIN outperforms the GIN on the vast majority of subjects and yields a more stable overall trend, suggesting that the proposed cross-subject supervised contrastive constraint can effectively suppress identity-related noise and improve the robustness of generalization.

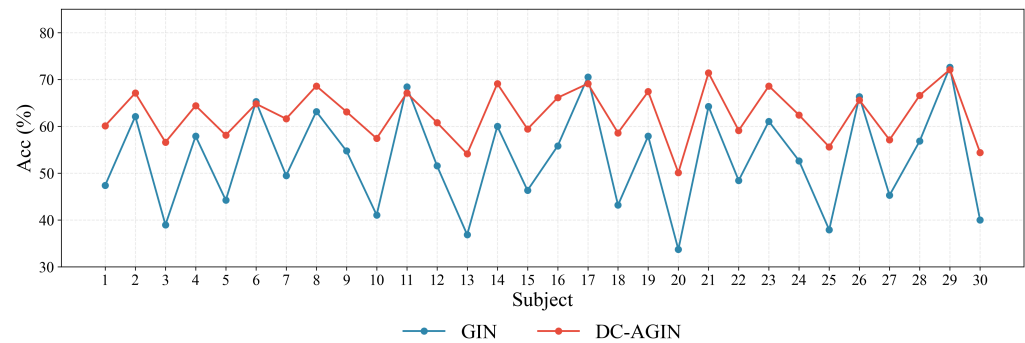


Figure 9. Per-subject subject-independent emotion recognition results of GIN and DC-AGIN.

To investigate the physiological basis of the model's decisions, we project the node-importance weights learned by the AGIN onto a two-dimensional channel-location distribution matrix. As shown in Figure 10, panel (a) provides the spatial position mapping matrix, and panel (b) visualizes the learned attention weights. The attention-weight visualization reflects a global pattern obtained by aggregating the attention weights of all test samples: we first compute the AGIN attention weight matrix for each sample and then average these matrices across all samples under the same setting to obtain a stable global attention distribution. The results clearly indicate that the prefrontal cortical region exhibits the most prominent activation (darkest color), while relatively high weights are also observed over the bilateral temporal areas. This distribution has clear neurophysiological implications: both the prefrontal and temporal lobes are known to be involved in emotion regulation and affect-related cognitive processing. These observations suggest that DC-AGIN can adaptively suppress background noise from less relevant regions and precisely focus on topologically informative nodes that are most discriminative for emotion recognition, thereby offering favorable neurophysiological interpretability.

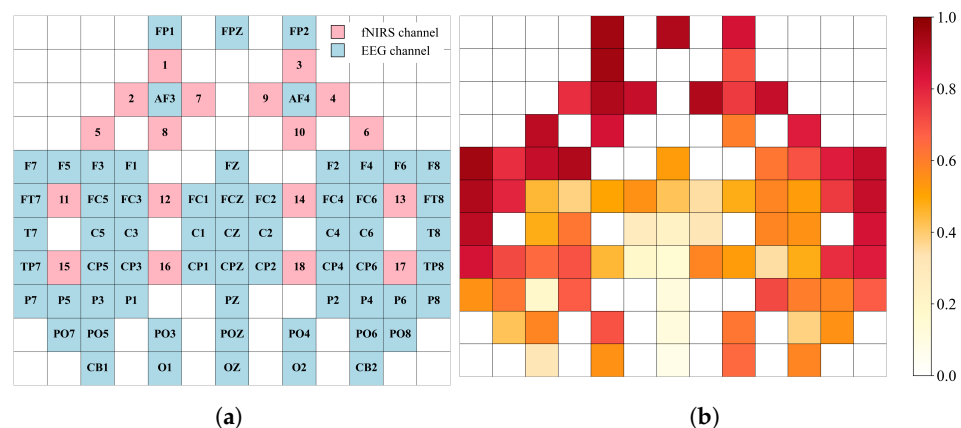


Figure 10. Attention visualization: (a) node spatial position mapping matrix; (b) attention map.

5. Conclusions

In this work, we propose DC-AGIN, a novel multi-modal fusion framework tailored for EEG–fNIRS emotion recognition. The framework integrates an attention-enhanced graph neural encoder with a dual-contrastive learning mechanism to address two key challenges in multi-modal cross-subject emotion recognition: modality heterogeneity and inter-subject variability. Specifically, the AGIN is first employed to adaptively aggregate topological features from key brain regions. Building upon this, cross-modal contrastive learning aligns the temporal dynamics of EEG with the spatially informative characteristics of fNIRS in the representation space. More importantly, we design a cross-subject supervised contrastive learning objective to explicitly suppress subject-identity noise, enabling the model to learn subject-invariant affective representations. Extensive experiments on the TYUT3.0 dataset demonstrate that DC-AGIN achieves strong accuracy and robustness under both five-fold subject-dependent validation and LOSO subject-independent evaluation, reaching state-of-the-art performance on this dataset.

In addition to its superior classification performance, this study provides strong evidence for the underlying neural mechanisms of emotion. This robustness, which is independent of specific subject calibration data, makes it a promising candidate for practical applications such as driver fatigue monitoring, mental stress assessment, and clinical diagnosis of affective disorders.

6. Discussion

Given the sensitivity of brain-signal data, this study places strong emphasis on data privacy and security and fully acknowledges the potential misuse risks of affective recognition technologies, such as unauthorized use, privacy infringement, and inappropriate decision support. Therefore, all data acquisition, storage, and processing in this work were conducted under strict ethical regulations and oversight, and the data were used solely for legitimate research purposes. Nevertheless, several limitations remain and warrant further investigation.

First, participants were primarily recruited among Chinese university students, leading to relatively homogeneous age and cultural backgrounds, which may limit the generalizability of the findings to broader populations. It should be noted that most publicly available EEG multi-modal emotion recognition datasets (e.g., SEED [31]) and related studies also predominantly involve university students or young adults, and their generalization ability may likewise be constrained. In addition, this study has not been independently validated on external public datasets; future work should incorporate more diverse samples and conduct systematic evaluations to more rigorously assess robustness and transferability.

Second, the EEG and fNIRS features used in this study are mainly segment-level static representations. While they can effectively support the current experimental paradigm and classification task, they do not explicitly model the temporal dynamics of affective states. Regarding feature construction, the fNIRS representation relies primarily on statistical descriptors, which may be insufficient to capture fine-grained hemodynamic variations and thus may limit the contribution of the fNIRS modality to affective discrimination. Future work will explore richer fNIRS features, such as time–frequency and wavelet-based dynamic descriptors, to enhance representational capacity. Moreover, incorporating dynamic feature extraction and temporal modeling frameworks may further improve the ability to characterize continuous affective changes in real-world scenarios.

In addition, the proposed framework has the potential to be extended to additional physiological modalities, such as ECG, GSR or EOG, which can provide complementary autonomic nervous system cues for affective states. When introducing new modalities,

future studies may follow the unified graph + contrastive learning paradigm adopted in this work: each modality can be encoded into modality-specific representations and aligned into a shared semantic space via cross-modal contrastive learning and/or consistency constraints. Furthermore, more advanced fusion strategies can be explored to achieve more thorough multi-modal integration, thereby forming a plug-and-play and extensible multi-physiological affective recognition framework.

Finally, although the current model exhibits the potential for real-time inference on GPUs, training remains computationally intensive, and its applicability across different devices, noise conditions, and clinical populations requires further validation. Future work will investigate model compression and acceleration to reduce training and deployment costs and to develop an efficient, low-latency online affective BCI system. In parallel, we will explore more advanced domain adaptation or adversarial techniques to further improve the model's adaptability in cross-subject and even cross-dataset affective recognition settings.

Author Contributions: Conceptualization, B.Y. and G.C.; methodology, B.Y.; software, B.Y.; validation, B.Y.; formal analysis, B.Y.; investigation, B.Y.; resources, X.Z.; data curation, X.Z.; writing—original draft preparation, B.Y.; writing—review and editing, B.Y., X.Z. and G.C.; visualization, B.Y.; supervision, X.Z. and G.C.; project administration, X.Z.; funding acquisition, X.Z. and G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research study was funded by the National Natural Science Foundation of China, grant numbers 62271342 and 62201377, and the Fundamental Research Program of Shanxi Province, China, grant number 202203021211174.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of Taiyuan University of Technology (TYUT-20220309, approved on 9 March 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the author due to privacy and ethical reasons. If you are interested in the dataset and want to use it, please download and fill out the license agreement at <https://gitee.com/tycgj/enter> (accessed on 1 December 2025) and send it to chengujun@tyut.edu.cn. We will send you a download link via email after reviewing your application. The model and training codes used in this study can be obtained through the following link: <https://github.com/CasiaYu/DC-AGIN> (accessed on 1 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Afzal, S.; Khan, H.A.; Piran, M.J.; Lee, J.W. A comprehensive survey on affective computing: Challenges, trends, applications, and future directions. *IEEE Access* **2024**, *12*, 96150–96168.
2. Mühl, C.; Allison, B.; Nijholt, A.; Chanel, G. A survey of affective brain computer interfaces: Principles, state-of-the-art, and challenges. *Brain-Comput. Interfaces* **2014**, *1*, 66–84.
3. Kaliouby, R.E.; Picard, R.; Baron-Cohen, S. Affective computing and autism. *Ann. N. Y. Acad. Sci.* **2006**, *1093*, 228–248.
4. Koelstra, S.; Mühl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31.
5. Susanto, I.Y.; Pan, T.Y.; Chen, C.W.; Hu, M.C.; Cheng, W.H. Emotion recognition from galvanic skin response signal based on deep hybrid neural networks. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 341–345.
6. Wang, F.; Mao, M.; Duan, L.; Huang, Y.; Li, Z.; Zhu, C. Intersession instability in fNIRS-based emotion recognition. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 1324–1333.
7. Ramadan, M.A.; Salem, N.M.; Mahmoud, L.N.; Sadek, I. Multimodal machine learning approach for emotion recognition using physiological signals. *Biomed. Signal Process. Control* **2024**, *96*, 166553.
8. Codina, T.; Blankertz, B.; Lüthmann, A.V. Multimodal fNIRS–EEG sensor fusion: Review of data-driven methods and perspective for naturalistic brain imaging. *Imaging Neurosci.* **2025**, *3*, IMAG-a.

9. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1263–1272.
10. Klepl, D.; Wu, M.; He, F. Graph neural network-based eeg classification: A survey. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2024**, *32*, 493–503.
11. Jia, J.; Zhang, B.; Lv, H.; Xu, Z.; Hu, S.; Li, H. CR-GCN: Channel-relationships-based graph convolutional network for EEG emotion recognition. *Brain Sci.* **2022**, *12*, 987.
12. Li, C.; Wang, F.; Zhao, Z.; Wang, H.; Schuller, B.W. Attention-based temporal graph representation learning for EEG-based emotion recognition. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 5755–5767.
13. Chen, G.; Liu, Y.; Zhang, X. EEG–fNIRS-Based Emotion Recognition Using Graph Convolution and Capsule Attention Network. *Brain Sci.* **2024**, *14*, 820.
14. Hou, M.; Zhang, X.; Chen, G.; Huang, L.; Sun, Y. Emotion Recognition Based on a EEG–fNIRS Hybrid Brain Network in the Source Space. *Brain Sci.* **2024**, *14*, 1166.
15. Chang, J.; Zhang, Z.; Qian, Y.; Lin, P. Multi-Scale Hyperbolic Contrastive Learning for Cross-Subject EEG Emotion Recognition. *IEEE Trans. Affect. Comput.* **2025**, *16*, 1716–1731.
16. Si, X.; He, H.; Yu, J.; Ming, D. Cross-subject emotion recognition brain–computer interface based on fNIRS and DBJNet. *Cyborg Bionic Syst.* **2023**, *4*, 45.
17. Li, X.; Song, D.; Zhang, P.; Zhang, Y.; Hou, Y.; Hu, B. Exploring EEG features in cross-subject emotion recognition. *Front. Neurosci.* **2018**, *12*, 162.
18. Deligani, R.J.; Borgheai, S.B.; McLinden, J.; Shahriari, Y. Multimodal fusion of EEG–fNIRS: A mutual information-based hybrid classification framework. *Biomed. Opt. Express* **2021**, *12*, 1635–1650.
19. Liu, C.; Zhou, X.; Wu, Y.; Yang, R.; Wang, Z.; Zhai, L.; Jia, Z.; Liu, Y. Graph neural networks in eeg-based emotion recognition: A survey. *arXiv* **2024**, arXiv:2402.01138.
20. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
21. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
22. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
23. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How powerful are graph neural networks? *arXiv* **2018**, arXiv:1810.00826.
24. Scholkmann, F.; Kleiser, S.; Metz, A.J.; Zimmermann, R.; Pavia, J.M.; Wolf, U.; Wolf, M. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *Neuroimage* **2014**, *85*, 6–27.
25. Duan, R.N.; Zhu, J.Y.; Lu, B.L. Differential entropy feature for EEG-based emotion classification. In Proceedings of the 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), San Diego, CA, USA, 6–8 November 2013; pp. 81–84.
26. Achard, S.; Bullmore, E. Efficiency and cost of economical brain functional networks. *PLoS Comput. Biol.* **2007**, *3*, e17.
27. Koszut, S.; Nallaperuma-Herzberg, S.; Lio, P. Decoding the Stressed Brain with Geometric Machine Learning. In *International Conference on AI in Healthcare*; Springer: Cham, Switzerland, 2025; pp. 3–16.
28. Shervashidze, N.; Schweitzer, P.; Van Leeuwen, E.J.; Mehlhorn, K.; Borgwardt, K.M. Weisfeiler-lehman graph kernels. *J. Mach. Learn. Res.* **2011**, *12*, 2539–2561.
29. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
30. Oord, A.V.D.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
31. Zheng, W.L.; Lu, B.L. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* **2015**, *7*, 162–175.
32. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
33. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
34. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing System, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
35. Zhang, J.; Wang, A.; Li, S.; Zhang, D.; Li, X. MFA-CNN: An Emotion Recognition Network Integrating 1D–2D Convolutional Neural Network and Cross-Modal Causal Features. *Brain Sci.* **2025**, *15*, 1165.
36. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 1180–1189.
37. Shen, X.; Liu, X.; Hu, X.; Zhang, D.; Song, S. Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2496–2511.

38. Wang, L.; Wang, S.; Jin, B.; Wei, X. GC-STCL: A Granger Causality-Based Spatial–Temporal Contrastive Learning Framework for EEG Emotion Recognition. *Entropy* **2024**, *26*, 540.
39. Gilakjani, S.S.; Al Osman, H. A graph neural network for EEG-based emotion recognition with contrastive learning and generative adversarial neural network data augmentation. *IEEE Access* **2023**, *12*, 113–130.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.