Knowledge-Based Systems 276 (2023) 110756

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

MTLFuseNet: A novel emotion recognition model based on deep latent feature fusion of EEG signals and multi-task learning

Rui Li, Chao Ren [*], Yiqing Ge, Qiqi Zhao, Yikun Yang, Yuhan Shi, Xiaowei Zhang, Bin Hu [*]

Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China

ARTICLE INFO

Article history:
Received 1 March 2023
Received in revised form 21 June 2023
Accepted 24 June 2023
Available online 30 June 2023

Keywords:
Emotion recognition
EEG
Feature fusion
Multi-task learning

ABSTRACT

How to extract discriminative latent feature representations from electroencephalography (EEG) signals and build a generalized model is a topic in EEG-based emotion recognition research. This study proposed a novel emotion recognition model based on deep latent feature fusion of EEG signals and multi-task learning, referred to as MTLFuseNet. MTLFuseNet learned spatio-temporal latent features of EEG in an unsupervised manner by a variational autoencoder (VAE) and learned the spatio-spectral features of EEG in a supervised manner by a graph convolutional network (GCN) and gated recurrent unit (GRU) network. Afterward, the two latent features were fused to form more complementary and discriminative spatio-temporal–spectral fusion features for EEG signal representation. In addition, MTLFuseNet was constructed based on multi-task learning. The focal loss was introduced to solve the problem of unbalanced sample classes in an emotional dataset, and the triplet-center loss was introduced to make the fused latent feature vectors more discriminative. Finally, a subject-independent leave-one-subject-out cross-validation strategy was used to validate extensively on two public datasets, DEAP and DREAMER. On the DEAP dataset, the average accuracies of valence and arousal are 71.33% and 73.28%, respectively. On the DREAMER dataset, the average accuracies of valence and arousal are 80.43% and 83.33%, respectively. The experimental results show that the proposed MTLFuseNet model achieves excellent recognition performance, outperforming the state-of-the-art methods.

1. Introduction

Emotion is the core component of human intelligence. The recognition and expression of emotion is essential for information exchange and understanding between people. Currently, emotion recognition has shown great potential in the field of artificial intelligence. Machines that can recognize human emotions will be more intelligent and efficient. For example, in health care, automatic emotion recognition can effectively aid in the diagnosis and treatment of mental disorders [1–3]. In the field of rehabilitation, recognizing emotions can also help to regulate and design rehabilitation exercises [4,5]. Since human emotional states are usually accompanied by physiological responses of the central nervous system and autonomic nervous system, researchers can use sensors to capture physiological signals to recognize emotions [6–8]. Electroencephalography (EEG) is an electrophysiological detection method that records the activity of the cerebral cortex through electrodes placed on the scalp. It can achieve more intuitive, reliable and objective recognition of human emotions. Recently, with the development of electrode and wearable technology, EEG-based emotion recognition has received extensive attention from researchers [9].

Various methods have been proposed to effectively analyze EEG emotional signals and realize emotion recognition. Typical machine learning methods mainly extract the time domain, frequency domain, time–frequency domain and nonlinear features of EEG signals [10,11] and then use support vector machine (SVM), k-nearest neighbors (KNN) and other classifiers to classify these features [12,13]. However, handcrafted feature extraction requires domain knowledge and expert experience [14] and may lose implicit information in raw EEG signals. With the ongoing development of deep learning technology, many researchers have considered building end-to-end deep learning models for EEG-based emotion recognition. These models can not only automatically extract complex features but also jointly optimize the learning and classification of EEG emotional features and achieve better emotion recognition results.

Although there have been many existing studies on EEG-based emotion recognition, EEGs are complex nonlinear signals, and the nonstationary property of EEG signals can easily lead to deviation in the distributions between different subjects [15]. In addition, EEG signals usually contain more intense noise than the image

[*] Corresponding authors.
E-mail addresses: ruili@lzu.edu.cn (R. Li), renc@lzu.edu.cn (C. Ren), 220220942860@lzu.edu.cn (Y. Ge), zhaoqq21@lzu.edu.cn (Q. Zhao), yangyk20@lzu.edu.cn (Y. Yang), 220220943121@lzu.edu.cn (Y. Shi), zhangxw@lzu.edu.cn (X. Zhang), bh@lzu.edu.cn (B. Hu).

https://doi.org/10.1016/j.knosys.2023.110756
0950-7051/© 2023 Elsevier B.V. All rights reserved.

signals that deep learning methods mainly deal with. Therefore, how to extract more discriminative and reliable features from EEG signals and establish a robust and universal deep learning emotion recognition model remains a challenge.

EEG signals have spatial, temporal and spectral patterns that are associated with a specific cognitive process [16]. By fusing the temporal, spatial and spectral features of EEG data, more complementary and discriminative emotional features can be obtained. Variational autoencoder (VAE) [17] is a kind of unsupervised learning model that is a deep generative network combining neural networks and probabilistic graphs. It is not only a generative model but also a feature extractor [18]. In this work, VAE was used to unsupervisedly learn low-dimensional latent representations from EEG data. Graph convolutional network (GCN) [19] provides a way to capture the intrinsic relationships between different nodes in the graph. It can be used to explore the implicit correlations between multiple graph nodes representing EEG channels. Gated recurrent unit (GRU) [20] is a recurrent neural network (RNN) for processing sequence data. It solves the long-term memory and gradient problems in RNN backpropagation and has lower computational cost. Based on the above models, this study proposed an EEG multidomain feature fusion method combining supervised and unsupervised learning to obtain more complementary and discriminative features of EEG signals. The processed EEG signals were fed into the VAE to learn spatio-temporal latent features. At the same time, GCN was used to learn spatial features, and then GRU was used to learn spectral domain features from the sequential list obtained by GCN to form spatio-spectral features. Then, the spatial, temporal, and spectral domain features of EEG signals were fused for emotion recognition.

Deep learning frameworks for emotion recognition suffer from poor generalization ability due to differences in EEG signals of different subjects. To address this issue, this study further introduced multi-task learning after completing the fusion of EEG spatial, temporal and spectral features. Multi-task learning trains multiple related tasks together so that different tasks can complement and promote each other to obtain better results than a single task. It is important to choose appropriate related tasks for effective multi-task learning [21]. The proposed multi-task model was trained for three emotion-related tasks: emotion recognition task, metric learning-based latent feature vector construction task and EEG data reconstruction task. The EEG data reconstruction task aimed to learn the latent feature vector of EEG signals, the emotion recognition task utilized EEG spatio-temporal–spectral fusion features to classify emotions, and the latent feature vector construction task further enhanced the separability of the fusion feature space. Additionally, to better classify the samples, focal loss [22] and triplet-center loss [23] were adopted in the emotion recognition task and latent feature vector construction task, respectively, and VAE loss was used in the EEG data reconstruction task. EEG emotion datasets have the problem of sample class imbalance, and focal loss can be used to solve the class imbalance problem in emotion recognition tasks. It reduces the weight of easy-to-classify samples and increases the weight of difficult-to-classify samples. Triplet-center loss is a metric loss function. It combines the advantages of triplet loss and center loss, which can effectively minimize the intraclass distance while maximizing the interclass distance of deep learning features.

This study proposed a novel emotion recognition model called MTLFuseNet based on deep latent spatio-temporal–spectral feature fusion of EEG signals and multi-task learning. The main contributions of this study are as follows:

1. A spatio-temporal–spectral EEG feature fusion method combining unsupervised and supervised learning is proposed. The spatio-temporal features and the spatio-spectral features of EEG are learned by VAE and GCN-GRU in an unsupervised and a supervised manner, respectively. Then, the two latent features are fused to form a more complementary data representation.
2. An emotion recognition model based on multi-task learning is proposed. The model consists of three tasks: emotion recognition, metric learning-based latent feature vector construction and EEG data reconstruction. To solve the problem of sample imbalance and to better classify the samples, focal loss and triplet-center loss are introduced.
3. An effective model with the most accurate and stable result is established. Extensive experiments have been conducted on two public datasets, DEAP [24] and DREAMER [25], and the results show that the proposed model can achieve state-of-the-art (SOTA) performance on both public datasets.

The rest of the paper is organized as follows: Section 2 introduces the existing EEG emotion recognition studies. Section 3 describes the definition of symbols. Section 4 presents a novel EEG-based emotion recognition model and describes its core components, including spatio-temporal latent feature extraction, spatio-spectral latent feature extraction, multi-task learning, and overall algorithm description. Section 5 describes a large number of experiments that were conducted on the DEAP and DREAMER datasets. Section 6 concludes this study.

2. Related work

2.1. EEG-based emotion recognition

A typical EEG emotion recognition process includes feature extraction and classification. The commonly used EEG features are statistics such as the mean value and variance, Hjorth parameters, power spectral density (PSD), wavelet energy, correlation dimension, Lyapunov exponent, sample entropy, differential entropy, etc. Jenke et al. [10] and Li et al. [15] reviewed these EEG features in the literature. After extracting the EEG features, the appropriate classifier is selected for emotion classification. SVM, KNN, naive Bayes, and random forest (RF) are commonly used classifiers. In addition, some researchers also use deep learning models as classifiers, which play the same role as traditional machine learning models. The choice of classifier depends on the extracted features and the specific emotion classification task, which can be one or more, or an ensemble learning method. For example, Yang et al. extracted Hjorth parameters, standard deviation, PSD, sample entropy, and wavelet entropy features from EEG signals and used SVM to achieve emotion recognition across subjects [26]. Rahman et al. employed Higuchi fractal dimension, sample entropy, permutation entropy features of EEG signals, and eight ensemble learning methods (RF, voting, bagging, AdaBoost, gradient boosting, XGBoost, light gradient boosting machine and stacking) to identify six basic emotions [27]. Thammasan et al. adopted a deep belief network to recognize music-emotion in a dynamic approach using EEG features fractal dimension, PSD, and discrete wavelet transform [28]. Anubhav et al. used band power as an EEG feature and long short-term memory (LSTM) as a classifier for emotion recognition. Compared to four classifiers, KNN, SVM, decision tree and RF, LSTM achieved the best performance [29].

Handcrafted feature extraction requires domain knowledge and is limited by human experience, which may lose latent information in raw EEG signals. In addition, the two stages of feature extraction and classification are separated, resulting in limited adaptability of the model. Therefore, an increasing number of end-to-end deep learning emotion recognition frameworks have been constructed to address this issue. For example, Wang et al. proposed a frame-level extraction neural network called FLDNet. FLDNet can capture EEG signals between different frame correlations and automatically learn to generate advanced features for emotion recognition [30]. Huang et al. developed a bihemisphere discrepancy convolutional neural network model (BiDCNN). BiDCNN enhanced the difference between the changes in the left and right cerebral hemispheres by mining the interchannel correlation between adjacent EEG electrodes, thus improving the effect of emotion recognition [31]. Guo et al. designed a horizontal and vertical feature fusion network based on different brain regions and hybrid dilation convolution for emotion recognition [32]. Liu et al. built an EEG emotion recognition model based on the attention mechanism and pretrained convolution capsule network to extract the global deep-seated features of EEG signals and obtain a better emotion recognition effect [33].

2.2. EEG feature representation

The EEG signal is time-varying and continuous, which has obvious time series characteristics. Neuroscience research has shown that human emotions are closely related to brain regions [34], and the correlation between channels representing the location of brain regions is an important indicator for emotion recognition. EEG signals contain rhythmic electrical activity, and many studies have also investigated the relationship between EEG spectra and human emotions [35,36]. Therefore, the emotion-related features of EEG signals exist in the temporal, spatial and spectral domains. Researchers usually extract temporal, spatial, or spectral features from raw EEG signals based on deep learning models for emotion recognition. For example, Alhagry et al. used LSTM to learn temporal features from EEG signals, and then the dense layer was used for emotion classification [37]. Wang et al. built a transformer-based EEG emotion recognition model to learn discriminative EEG spatial information from the electrode level to the brain-region level [38]. Demir et al. constructed an EEG emotion classification framework for the automatic classification of human emotion using EEG spectral information. This framework employed wavelet transform and continuous wavelet transform to convert the EEG signals into EEG rhythm images, and then five well-known pretrained CNN models, AlexNet, VGG16, ResNet50, SqueezeNet and MobilNetv2, were used for feature extraction [39].

In recent works, to obtain better EEG feature representation, researchers have begun to fuse multidomain EEG features to obtain more complementary and discriminative emotional features. For example, Liu et al. designed a deep learning emotion recognition model that was composed of a spatio-temporal feature extraction module and an EEG channel attention weight learning module to extract the discriminant features of multichannel EEG signals during a continuous period of time [40]. Miao et al. developed a multiple frequency band parallel spatial–temporal 3D deep residual learning framework for EEG-based emotion recognition [41]. Feng et al. established a hybrid deep learning model that includes a spatial-graph convolutional network module and an attention-enhanced BiLSTM memory module. It can consider the biological topology information of each brain region to extract representative spatial–temporal features from multiple EEG channels [42]. Yao et al. proposed a 3D feature construction method based on the fusion of EEG spatial and spectral information and constructed a deep learning classification framework based on feature fusion representation and a dilated bottleneck-based convolutional network for emotion recognition [43]. Li et al. presented a novel spatial-frequency convolutional self-attention network combining feature learning in the spatial and spectral domains of EEG signals for emotion recognition. It employed a parallel convolutional neural network layer to capture the spatial information, and then the embedded intrafrequency band self-attention was used to learn the spectral information [44].

All the above studies are based on spatio-temporal or spatio-spectral EEG feature fusion, and feature extraction and fusion are performed by supervised learning, which may ignore latent EEG features. Currently, unsupervised learning has been used by researchers to explore hidden structures in EEG data in a more natural way to obtain latent features [45,46]. How to effectively integrate supervised learning and unsupervised learning and use them for EEG signal feature extraction and model construction is a topic worth researching. Therefore, this study combined supervised learning and unsupervised learning to extract and integrate spatial, temporal and spectral domain features of EEG signals to form a more effective model for emotion recognition.

2.3. Multi-task learning

Multi-task learning [47] is a machine learning method that learns at least two tasks at the same time. By sharing the correlation between multiple tasks, it can avoid underfitting training and improve the generalization ability of the model. Multi-task learning has been successfully applied in natural language processing, speech recognition, computer vision and many other fields [48–51]. Most of the existing EEG-based emotion recognition studies are single-task, and some researchers consider introducing multi-task learning into this field. For example, Li et al. developed a novel model for EEG-based emotion recognition based on multi-task learning with a capsule network and attention mechanism. This model can share complementary information from the arousal, valence and dominance tasks, thus capturing more information from the EEG signals of these three tasks [52]. Priyasad et al. presented EEG channelwise encoder networks coupled with geometric deep learning and multi-task learning for emotion recognition, exploring the viability of multi-task learning to train classifiers to simultaneously recognize the three emotional dimensions: arousal, valence and dominance [53]. Li et al. designed a graph-based multi-task self-supervised learning model (GMSS) for EEG emotion recognition. The GMSS has three tasks, including spatial and frequency jigsaw puzzle tasks and contrastive learning task, and it can integrate multiple tasks and capture a more general representation of all tasks [54]. Therefore, this study introduced multi-task learning after fusing the EEG spatial, temporal and spectral features. By combining three emotion-related tasks, more general EEG representations were learned to improve the generalization performance of the model.

3. Preliminaries

In this study, the raw EEG signals for a trial can be defined as follows:

$$X = (x_1, x_2, \ldots, x_C) \in R^{C\times L}, \tag{1}$$

where the trial collected EEG data on C channels that lasted for T seconds, $L = T \times sr$, sr represents the sampling rate. The raw EEG data collected on the ith channel can be formulated as follows:

$$x_i = (s_1, s_2, \ldots, s_T) \in R^{T\times sr}(i \in \{1, 2, \ldots, C\}), \tag{2}$$

where $s_t \in R^{sr}(t \in \{1, 2, \ldots, T\})$ represents EEG data acquired on a single channel in 1 s. The raw EEG signals pass through a custom finite impulse response (FIR) bandpass filter with a Hanning window to extract five frequency bands: δ, θ, α, β and γ. Because the δ frequency band is generally considered to be exclusively associated with sleep, it was excluded from this study. EEG signals on different frequency bands can be expressed as follows:

$$X_B = (x'_1, x'_2, \ldots, x'_C) \in R^{C\times L}, B \in \{\theta, \alpha, \beta, \gamma\}. \tag{3}$$

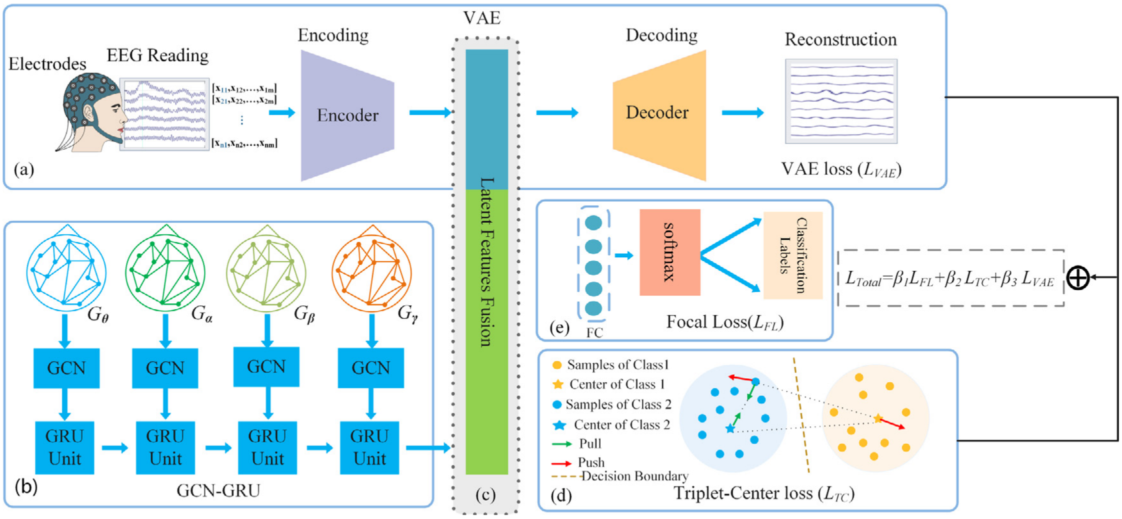


Fig. 1. The flowchart of the MTLFuseNet model: (a) the spatio-temporal latent feature extraction process based on VAE; (b) the spatio-spectral latent feature extraction process based on GCN-GRU; (c) latent feature fusion process; (d) (e) loss function construction and emotion classification.

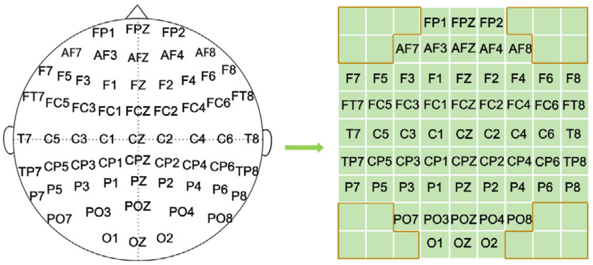

FP1 FPZ FP2
AF7 AF3 AFZ AF4 AF8
F7 F5 F3 F1 FZ F2 F4 F6 F8
FT7 FC5 FC3 FC1 FCZ FC2 FC4 FC6 FT8
T7 C5 C3 C1 CZ C2 C4 C6 T8
TP7 CP5 CP3 CP1 CPZ CP2 CP4 CP6 TP8
P7 P5 P3 P1 PZ P2 P4 P6 P8
PO7 PO3 POZ PO4 PO8
O1 OZ O2
FP1 FPZ FP2
AF7 AF3 AFZ AF4 AF8
F7 F5 F3 F1 FZ F2 F4 F6 F8
FT7 FC5 FC3 FC1 FCZ FC2 FC4 FC6 FT8
T7 C5 C3 C1 CZ C2 C4 C6 T8
TP7 CP5 CP3 CP1 CPZ CP2 CP4 CP6 TP8
P7 P5 P3 P1 PZ P2 P4 P6 P8
PO7 PO3 POZ PO4 PO8
O1 OZ O2

Fig. 2. The flowchart of EEG signals converted into a matrix.

4. Methodology

This study proposed a novel emotion recognition model based on deep latent feature fusion of EEG signals and multi-task learning, called MTLFuseNet. Fig. 1 shows the overall framework of the proposed model. This part is composed of the following parts: spatio-temporal latent feature extraction based on VAE, spatio-spectral feature extraction based on GCN and GRU, loss function definition in multi-task learning, and overall algorithm description.

4.1. Spatio-temporal latent feature extraction based on VAE

4.1.1. Spatio-temporal encoding

At present, the electrode placement rules in EEG signal acquisition are generally based on the international 10–20 system. To further characterize spatial relationships, EEG channels were encoded based on the position of the electrodes. As shown in Fig. 2, the electrodes are converted into a matrix representation of $H \times W$. Usually, H and W are set to 9, and empty areas marked in red are usually set to 0. The spatio-temporal encoding of the raw EEG signals for a trial can be defined as follows:

$$X^{ST} \in R^{L \times H \times W}. \tag{4}$$

4.1.2. Variational autoencoder (VAE)

VAE [17] provides a closed-form solution of the underlying distribution of the input data. It is an unsupervised learning method suitable for the extraction of latent features of EEG signals. The extracted latent features can be used as low-dimensional and robust feature representations of EEG signals. Similar to the standard autoencoder (AE), the VAE consists of two parts: an encoder and a decoder. VAE learns the probability distribution parameters of the latent space representation Z_{ST} of EEG spatio-temporal encoded data X^{ST}. The true distribution of the input data X^{ST} can be defined as $P_\theta(X^{ST})$:

$$P_\theta(X^{ST}) = \int_{Z_{ST}} P_\theta(X^{ST}|Z_{ST}) P_\theta(Z_{ST})\, dZ_{ST}. \tag{5}$$

Since the above formula involves the integration of high-dimensional variables, this is difficult to solve. It is also difficult to solve by posterior distribution, $P_\theta(Z_{ST}|X^{ST}) = \frac{P_\theta(X^{ST}|Z_{ST})P_\theta(Z_{ST})}{P_\theta(X^{ST})}$, and $P_\theta(X^{ST})$ is involved again. To speed up the calculus to make it feasible, it is necessary to introduce a further function to approximate the posterior distribution $Q_\phi(Z_{ST}|X^{ST})$:

$$Q_\phi(Z_{ST}|X^{ST}) \approx P_\theta(Z_{ST}|X^{ST}). \tag{6}$$

As shown below, the VAE measures the distance between two distributions by KL divergence and minimizes the distance by optimization methods.

$$minD_{KL}\left(p_\theta\left(Z_{ST} \mid X^{ST}\right) || Q_\phi(Z_{ST}|X^{ST})\right). \tag{7}$$

By deriving the above equation, we can obtain the maximization problem of the following equation:

$$\begin{aligned} \pounds_{VAE} = & \mathbb{E}_{Q_\phi(Z_{ST}|X^{ST})}[\log P_\theta(X^{ST}|Z_{ST})] \\ & - D_{KL}(Q_\phi(Z_{ST}|X^{ST})||P_\theta(Z_{ST})). \end{aligned} \tag{8}$$

This equation is called the variational lower bound, the first equation represents the possibility of reconstructing likelihood, and the second equation ensures that the distribution of neural network learning $Q_\phi(Z_{ST}|X^{ST})$ is similar to the prior distribution of the real latent feature $P_\theta(Z_{ST})$. Assuming $Q_\phi(Z_{ST}|X^{ST})$ conforms to the Gaussian distribution, using a multivariate Gaussian distribution with a diagonal covariance structure, that is, $logQ_\phi(Z_{ST}|X^{ST}) = log\mathcal{N}(Z_{ST}; \mu, \sigma^2 I)$, where I is the diagonal matrix, and μ and σ^2 represent the mean and variance of the approximate posterior probability, respectively. Since the sampling operation is nondifferentiable and cannot be backpropagated, this problem can be solved by the reparameterization trick. For the data x_i^{ST}, the variational lower bound can be converted into the

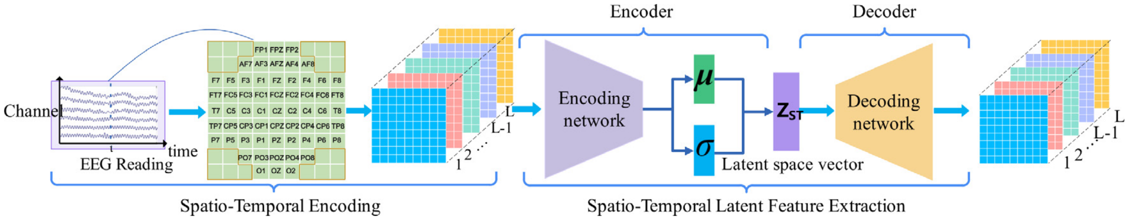

Encoder
Decoder
Channel
time
EEG Reading
FP1 FPZ FP2
AF7 AF3 AFZ AF4 AF8
F7 F5 F3 F1 FZ F2 F4 F6 F8
FT7 FC5 FC3 FC1 FCZ FC2 FC4 FC6 FT8
T7 C5 C3 C1 CZ C2 C4 C6 T8
TP7 CP5 CP3 CP1 CPZ CP2 CP4 CP6 TP8
P7 P5 P3 P1 PZ P2 P4 P6 P8
PO7 PO3 POZ PO4 PO8
O1 OZ O2
L
L-1
2
1
Encoding network
μ
σ
Z_ST
Decoding network
Latent space vector
Spatio-Temporal Encoding
Spatio-Temporal Latent Feature Extraction

Fig. 3. Flowchart of spatio-temporal latent feature extraction.

following form after the above operations:

$$\pounds_{VAE} \simeq \frac{1}{2}\sum_{j=1}^{J}\left(1+\log\left(\sigma_j^{(i)2}\right)-\mu_j^{(i)2}-\sigma_j^{(i)2}\right) + \frac{1}{N}\sum_{n=1}^{N}\log p_\theta\left(x_i^{ST} \mid z_{ST}^{(i,n)}\right), \tag{9}$$

where the latent variable Z_{ST} is rewritten by adding noise variable ε and distribution conversion function g_ϕ. $z_{ST}^{(i,n)} = g_\phi(x_i^{ST}, \varepsilon^{(n)}) = \mu^{(i)} + \sigma^{(i)} \odot \varepsilon^{(n)}$, $\varepsilon^{(n)} \sim \mathcal{N}(0, I)$, J is the dimension of Z_{ST}, N is the number of samples, and $\odot$ is the product of elements.

4.1.3. Spatio-temporal latent feature extraction process

As shown in Fig. 3, the spatio-temporal latent feature extraction process based on VAE includes two stages: spatio-temporal data encoding and spatio-temporal latent feature extraction. First, the original EEG signal was encoded into spatio-temporal data according to Eq. (4), and the encoded data X^{ST} were obtained. After that, X^{ST} was used as input data and fed into the VAE to obtain the spatio-temporal latent features Z_{ST}.

4.2. Spatio-spectral latent feature extraction based on GCN and GRU

4.2.1. Spatio-spectral graph sequence construction

Definition 1. Spatio-Spectral Graph Representation. In this study, the graph used is defined as $\mathcal{G} = (V, E, A)$, where V represents the set of nodes, E represents the set of edges and $A = (a_{1,1}, \dots, a_{C,C}) \in R^{C\times C}$ is the adjacency matrix. $|V| = C$, C is the number of channels. For a given pair of channels (x_i, x_j), the correlation between them can be defined as follows:

$$a_{i,j} = I\left(x_i, x_j\right) = \sum_{u\in x_i}\sum_{v\in x_j} p(u,v)\log\frac{p(u,v)}{p(u)p(v)}, \tag{10}$$

where $1 \leqslant i \leqslant C, 1 \leqslant j \leqslant C$, x_i and x_j are defined according to Eq. (2). In this study, the differential entropy (DE) feature of each channel extracted from different frequency bands B and the adjoining matrix defined above construct the spatio-spectral graph $\mathcal{G}_B = (V', E', A)$, $B \in \{\theta, \alpha, \beta, \gamma\}$.

Definition 2. Spatio-Spectral Graph Sequence. In this study, the spatio-spectral graph sequence is constructed based on the spatio-spectral graph on each frequency band as follows:

$$\mathcal{G}^S = (\mathcal{G}_\theta, \mathcal{G}_\alpha, \mathcal{G}_\beta, \mathcal{G}_\gamma). \tag{11}$$

4.2.2. Graph convolutional networks (GCN)

There are many functional connections between different regions of the brain, and analyzing the spatial relationship between these connections is helpful for emotion recognition tasks. GCN [19] is based on graph theory, which uses spectral graph theory to realize the convolutional structure and extends the convolutional operation from image-like data to graph structure data. In this study, GCN was used to capture the correlation between different EEG channels.

In spectrogram analysis, the graph structure is represented by the corresponding Laplacian matrix. The properties of the graph structure are obtained by analyzing the Laplacian matrix and its eigenvalues. The Laplacian matrix is defined as $LA = D - A$, $LA \in R^{C\times C}$, where $A \in R^{C\times C}$ is an adjacency matrix, and its definition refers to Definition 1. $D \in R^{C\times C}$ is the degree matrix of graph nodes. The Laplacian matrix after symmetric normalization can be expressed as $LA = I_N - D^{-\frac{1}{2}}AD^{\frac{1}{2}}$, and the eigenmatrix $LA = U_L\Lambda_L U_L^T$ can be obtained by eigendecomposition, where I_N is the identity matrix and U_L is the basis of the Fourier transform, that is, the eigenvector matrix of LA; $\Lambda_L = diag([\lambda_1, \dots, \lambda_C]) \in R^{C\times C}$ is a diagonal matrix of eigenvalues. In this study, the node feature of graph $\mathcal{G}_B$ in graph sequence $\mathcal{G}^S$ is defined as $V'_B \in R^C$, and the Fourier transform of the graph can be represented as $\hat{V'_B} = U_L^T V'_B$. According to the property of the Laplacian matrix, U_L is an orthogonal matrix, and the corresponding inverse Fourier transform is $V'_B = U_L\hat{V'_B}$. The convolution operation of node feature V'_B on graph $\mathcal{G}_B$ can be defined as:

$$g_\theta \bigstar_{\mathcal{G}_B} V'_B = U_L g_\theta U_L^T V'_B, \tag{12}$$

where g_θ represents the graph convolution kernel, and $\bigstar_{\mathcal{G}_B}$ is the convolution operation on the spatio-spectral graph $\mathcal{G}_B$.

Since it is expensive to compute the eigendecomposition of the Laplacian matrix LA of the large graph, the Chebyshev polynomial $T_k(\tilde{LA})$ can be used to approximate the convolution kernel and reduce the computational complexity. Convolution can be redefined as follows:

$$g_\theta \bigstar_{\mathcal{G}_B} V'_B \approx \sum_{k=0}^{K}\theta_k T_k(\tilde{LA})V'_B, \tag{13}$$

where $\tilde{LA} = \frac{2}{\lambda_{max}}LA - I_N$, λ_{max} is the largest eigenvalue of the Laplacian matrix LA, θ_k is the polynomial coefficient, and θ is the polynomial coefficient vector. Chebyshev polynomials can be recursively defined as $T_K(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, where $T_0(x) = 1$, $T_1(x) = x$. Based on the above approximate calculation, spectral convolution no longer depends on the whole graph but only on nodes within steps from the central node K.

For fast calculation, $\lambda_{max} \approx 2$, $K = 1$ can be set, and the graph convolution can be further simplified into the following formula:

$$g_\theta \bigstar_{\mathcal{G}_B} V'_B \approx \theta(I_N + D^{-\frac{1}{2}}AD^{\frac{1}{2}})V'_B. \tag{14}$$

It is worth noting that $I_N + D^{-\frac{1}{2}}AD^{\frac{1}{2}}$ has a value range of [0,2], which is multiplied repeatedly when used, resulting in numerical instability and gradient explosion/disappearance. To alleviate this problem, the following reregularization techniques are used: $I_N + D^{-\frac{1}{2}}AD^{\frac{1}{2}} \rightarrow \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{\frac{1}{2}}$, where $\tilde{A} = A + I_N$, $\tilde{D}_{ii} = \Sigma_j\tilde{A}_{ij}$. Thus, the graph convolution operation of graph $\mathcal{G}_B$ on the spatio-spectral graph sequence $\mathcal{G}^S$ can be defined as:

$$H_B^{GCN} = \sigma_{re}(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{\frac{1}{2}}V'_B W + b), \tag{15}$$

where H_B^{GCN} represents the feature vector of the EEG signal after graph convolution, W is the weight parameter, b is the bias parameter, and σ_{re} represents the ReLU activation function.

Applying the above graph convolution operation to each spatio-spectral graph $\mathcal{G}_B$ in the spatio-spectral graph sequence $\mathcal{G}^S$ can obtain a sequence containing the spatial information of EEG signals $H^{GCN^S} = \{H_\theta^{GCN}, H_\alpha^{GCN}, H_\beta^{GCN}, H_\gamma^{GCN}\}$. To reduce the number of parameters, all graph convolution operations on spatio-spectral graph sequence share the same parameters.

4.2.3. Gated recurrent unit (GRU)

GRU [20] is considered a variant of LSTM and is designed to solve the gradient problem introduced by standard recurrent neural networks. GRU lacks the output gate compared to LSTM, and it is composed of an update gate and a reset gate, so GRU needs fewer parameters. In this study, sequence H^{GCN^S} extracted from the GCN network containing EEG spatial information was used as the input of the GRU. GRU was used to further capture the correlation between different frequency bands of EEG signals to obtain spatio-spectral latent features of EEG signals Z_{SS}.

For the GCN output sequence H^{GCN^S}, the hidden state on the Bth frequency band is calculated as follows:

$$H_B = (1 - Z_B) \odot H_{B-1} + Z_B \odot \tilde{H}_B, \tag{16}$$

where Z_B is the update gate vector, H_{B-1} is the previous state, and $\tilde{H}_B$ is the candidate activation vector. The operator $\odot$ denotes the Hadamard product.

The update gate vector Z_B is calculated as follows:

$$Z_B = \sigma_g(W_z H_B^{GCN} + V_z H_{B-1} + b_z), \tag{17}$$

where σ_g is a sigmoid function, H_B^{GCN} is the input vector on frequency band B; W_z, V_z and b_z are parameter matrices and vectors, and H_{B-1} is the hidden state of the $B-1$ frequency band.

The candidate activation vector $\tilde{H}_B$ is calculated as follows:

$$\tilde{H}_B = \sigma_h(W_h H_B^{GCN} + V_h(R_B \odot H_{B-1}) + b_h), \tag{18}$$

where σ_h is a hyperbolic tangent, W_h and V_h are the weight parameters, b_h is the bias, and R_B is the reset gate that controls the contribution of the previous state to the current candidate state $\tilde{H}_B$.

The reset gate R_B is calculated as follows:

$$R_B = \sigma_g(W_r H_B^{GCN} + V_r H_{B-1} + b_r), \tag{19}$$

where W_r and V_r are weight parameters and b_r is the bias.

4.2.4. Spatio-spectral latent feature extraction process

As shown in Fig. 4, the extraction process of the spatio-spectral latent feature based on GCN-GRU includes two stages: the construction of the spatio-spectral graph sequence and the extraction of the spatio-spectral latent feature. The spatio-spectral graph sequence constructed by the original EEG signal was input into the GCN network to obtain the spatial correlation information between the EEG channels. Then, the sequence H^{GCN^S} output by the GCN network containing the spatial information of EEG was sent to the GRU network to further obtain the correlation between different EEG frequency bands. The final output Z_{SS} of the GRU network is the spatio-spectral latent feature of the EEG signal.

4.3. Multi-task learning

To improve the accuracy and generalization ability of the emotion recognition model, multi-task learning was introduced in this study. Multi-task learning trains multiple related tasks together, making different tasks complement and promote each other, thus obtaining better results than single-task learning. The multi-task learning model proposed in this study has three emotion-related tasks: emotion recognition, metric learning-based latent feature vector construction and EEG data reconstruction. To better classify the samples, focal loss and triplet-center loss were used in the emotion recognition task and latent feature vector construction task, respectively. The loss function in the reconstruction task used the VAE loss, as shown in Eq. (9).

4.3.1. Focal loss

EEG emotion datasets have the problem of sample class imbalance, and focal loss can be used to solve the class imbalance problem in emotion classification tasks. Focal loss was first proposed by Lin et al. [22] to address the data imbalance problem in object detection. The focal loss function is an improvement on the standard cross entropy loss. The standard cross entropy loss does not address the importance of positive/negative samples, but the focal loss function allows the model to be trained to focus more on hard-to-classify samples by reducing the weight of easily classified samples. The focal loss is defined as follows:

$$\pounds_{FL} = \begin{cases} -\alpha(1-\hat{y})^\gamma \log \hat{y}, \, if \;\; y = 1 \\ -(1-\alpha)\hat{y}^\gamma \log(1-\hat{y}), \, if \;\; y = 0 \end{cases} \tag{20}$$

where α is the balance factor, γ is used to adjust the rate at which the weights of simple samples are reduced, and $\hat{y}$ represents the output of the activated function.

4.3.2. Triplet-center loss

Deep metric learning is a spatial mapping method that obtains a more separable feature space by learning the degree of similarity between data so that similar sample feature vectors are closer and dissimilar sample feature vectors are farther away. The loss function plays an important role in deep metric learning. Typical metric loss functions include contractive loss, triplet loss and center loss.

Triplet-center loss [23] fully combines the advantages of triplet loss and center loss, and its goal is to effectively minimize the intraclass distance in feature space and maximize the interclass distance in feature space, as shown in Fig. 1(d). For M minibatch training samples, the triplet-center loss is defined as follows:

$$\pounds_{TC} = \sum_{i=1}^{M} \max\left(D\left(f_i, c_{y^i}\right) + m - \min_{j \neq y^i} D\left(f_i, c_j\right), 0\right), \tag{21}$$

where m represents the margin, y^i represents the data label of the ith sample, c_j represents the center of the sample class labeled j, f_i represents the vector representation of the ith sample through the neural network f, and $D(\cdot)$ represents the square Euclidean distance function, which is defined as $D(f_i, c_{y^i}) = \frac{1}{2}||f_i - c_{y^i}||_2^2$.

4.3.3. Total loss

In this study, the training objective optimization function of the model is a combination of the three loss functions: focal loss ($\pounds_{FL}$), triplet-center loss ($\pounds_{TC}$) and VAE loss ($\pounds_{VAE}$). The final loss function of the proposed MTLFuseNet model $\pounds_{Total}$ is defined as follows:

$$\pounds_{Total} = \beta_1 \pounds_{FL} + \beta_2 \pounds_{TC} + \beta_3 \pounds_{VAE}, \tag{22}$$

where β_1, β_2, and β_3 are hyperparameters to the degree of contribution of each loss function.

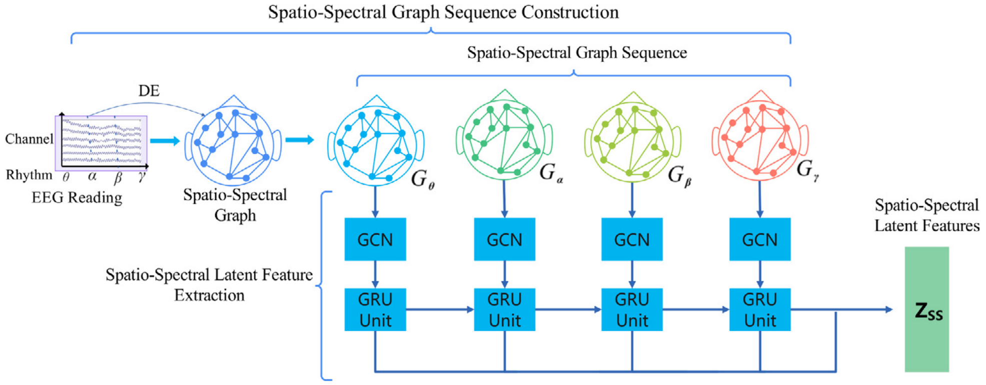


Fig. 4. Flowchart of spatio-spectral latent feature extraction.

4.4. MTLFuseNet for EEG-based emotion recognition

To extract the potential features of EEG signals from different perspectives, a deep latent feature fusion emotion recognition model based on multi-task learning was constructed in this study, which is called the MTLFuseNet model.

First, spatio-temporal data were encoded on the raw EEG signals and fed into the VAE network to learn the spatio-temporal latent features of the EEG signals. Both the encoder and decoder of the VAE are composed of four convolutional layers. Fig. 3 and Fig. 1(a) show in detail the process of extracting spatio-temporal latent features Z_{ST} with the VAE network. At the same time, to better extract the spatio-spectral information contained in EEG signals, the spatio-spectral graph sequence constructed by the original EEG signal was input into the GCN network to obtain the spatial correlation information between the EEG channels. The output of the GCN was used as the input of the GRU network, and the spatio-spectral latent features of EEG signals were obtained by the GRU network. Fig. 4 and Fig. 1(b) show the process of extracting spatio-spectral latent features Z_{SS} using the GCN-GRU network. Finally, the spatio-temporal latent features Z_{ST} and the spatio-spectral latent features Z_{SS} were fused to form the spatio-temporal–spectral fusion features. The process of feature fusion is shown in Fig. 1(c). The fusion operation is shown in Eq. (23):

$$Z_{SST} = [Z_{ST} \oplus Z_{SS}] , \tag{23}$$

where $\oplus$ is the concatenate operation.

In addition, MTLFuseNet was constructed based on multi-task learning. Considering the problem of sample category imbalance in emotion datasets, focal loss was used in the emotion classification task. In addition, triplet-center loss was introduced to make the fused latent feature vectors more discriminative to obtain a more separable feature space. $\pounds_{FL}$, $\pounds_{TC}$, and $\pounds_{VAE}$ were fused to construct the final loss function $\pounds_{Total}$, as shown in Fig. 1(d) and (e).

To realize network joint training, algorithm 1 shows the joint optimization algorithm of the MTLFuseNet model. In each iteration, Z_{ST}, Z_{SS} and the network parameters of the VAE network, GCN network and GRU network are updated. The weight parameters and bias of the fully connected (FC) layer are updated after each iteration. After the FC layer, the predicted value $\hat{Y}$ is calculated according to Eq. (24). In the algorithm 1, $\left[Z_{ST}^t \oplus Z_{SS}^t\right]$ represents the fusion feature vector, where Z_{ST}^t represents the output of the VAE network in the tth iteration. Z_{SS}^t represents the output of the GCN-GRU network in the tth iteration. The MTLFuseNet model uses the $\pounds_{Total}$ loss as the loss function.

$$\hat{Y} = softmax(\omega_F Z_{SST} + b_F). \tag{24}$$

Algorithm 1 MTLFuseNet joint optimization algorithm

```
Require: A set of EEG training samples ($X^{ST}$, $\mathcal{G}^S$, $Y$), and $Y$ are the training labels;
Ensure: parameters of VAE: $\omega_v$; parameters of GCN: $\omega_{gcn}$; parameters of GRU: $\omega_{gru}$; parameters of FC layer: $\omega_F$
initial $t = 1$ ; $MaxIter$;
for $t \in [1, MaxIter]$ do
    $Z_{ST}^{t+1} = VAE\left(X^{ST}, \omega_v^t\right)$
    $H^{GCN^S\,t+1} = GCN(\mathcal{G}^S, \omega_{gcn}^t)$
    $Z_{SS}^{t+1} = GRU(H^{GCN^S}, \omega_{gru}^t)$
    $Z_{SST}^{t+1} = \left[Z_{ST}^{t+1} \oplus Z_{SS}^{t+1}\right]$
    $\hat{Y} = F\left(Z_{SST}^{t+1}, \omega_F^t\right)$
    $\nabla\omega_v^{t+1} = \frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z_{ST}^{t+1}} \cdot \frac{\partial Z_{ST}^{t+1}}{\partial \omega_v^t}$
    $\nabla\omega_{gru}^{t+1} = \frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z_{SS}^{t+1}} \cdot \frac{\partial Z_{SS}^{t+1}}{\partial \omega_{gru}^t}$
    $\nabla\omega_{gcn}^{t+1} = \frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z_{SS}^{t+1}} \cdot \frac{\partial Z_{SS}^{t+1}}{\partial H^{GCN^S\,t+1}} \cdot \frac{\partial H^{GCN^S\,t+1}}{\partial \omega_{gcn}^t}$
    $\widetilde{\omega}_v^{t+1} = \omega_v^t - \eta_{vae}\nabla\omega_v^{t+1}$
    $\widetilde{\omega}_{gru}^{t+1} = \omega_{gru}^t - \eta_{gru}\nabla\omega_{gru}^{t+1}$
    $\widetilde{\omega}_{gcn}^{t+1} = \omega_{gcn}^t - \eta_{gcn}\nabla\omega_{gcn}^{t+1}$
    $\nabla\omega_F^{t+1} = \frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial \omega_F}$
    $\widetilde{\omega}_F^{t+1} = \omega_F^t - \eta_F\nabla\omega_F^{t+1}$
end for
```

In the reverse propagation process of the MTLFuseNet model, the synchronization of spatio-temporal and spatio-spectral networks is involved. For the update of the spatio-temporal subnetwork, first, the gradient $\frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}}$ of the FC layer is solved, then the gradient $\frac{\partial \hat{Y}}{\partial Z_{ST}^{t+1}}$ of the spatio-temporal features applied to the VAE network by the FC layer is calculated, and finally, the gradient $\frac{\partial Z_{ST}^{t+1}}{\partial \omega_v^t}$ of the parameters of the whole VAE network is calculated. According to the chain rule, the VAE network parameters and the gradient $\nabla\omega_v^{t+1} = \frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z_{ST}^{t+1}} \cdot \frac{\partial Z_{ST}^{t+1}}{\partial \omega_v^t}$ corresponding to the parameters of the FC layer can be obtained, which is the process of solving the gradient of the parameters of the spatio-temporal subnetwork. Then, the parameters of the VAE subnetwork are updated according to $\widetilde{\omega}_v^{t+1} = \omega_v^t - \eta_{vae}\nabla\omega_v^{t+1}$, and the parameters of the spatio-temporal subnetwork are updated once. For the updating process of the spatio-spectral subnetwork, first, the gradient $\frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}}$ of the parameters of the FC layer of the network is solved; second, the gradient $\frac{\partial \hat{Y}}{\partial Z_{SS}^{t+1}}$ of the output implicit features of the GRU network acted by the FC layer is calculated; then, the gradient $\frac{\partial Z_{SS}^{t+1}}{\partial \omega_{gru}^{t}}$ of the parameters of the whole GRU network is calculated. According to the chain rule, the GRU network parameters and the gradient $\nabla \omega_{gru}^{t+1} = \frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z_{SS}^{t+1}} \cdot \frac{\partial Z_{SS}^{t+1}}{\partial \omega_{gru}^{t}}$ corresponding to the parameters of the FC layer can be obtained. Here is the gradient solving process of GRU subnetwork parameters, and the parameters of the GRU subnetwork are updated according to $\widetilde{\omega}_{gru}^{t+1} = \omega_{gru}^{t} - \eta_{gru} \nabla \omega_{gru}^{t+1}$. After that, the network parameters of the GCN in the spatio-spectral subnetwork are further updated. Similar to the updating process of the GRU, it is necessary to calculate the gradient $\frac{\partial Z_{SS}^{t+1}}{\partial {H^{GCN}}^{S\,t+1}}$ of the output implicit features of the GCN network by the GRU network and the gradient $\frac{\partial {H^{GCN}}^{S\,t+1}}{\partial \omega_{gcn}^{t}}$ of the entire GCN network parameters. According to the chain rule, the parameters of the GCN network and the gradient $\nabla \omega_{gcn}^{t+1} = \frac{\partial \pounds_{Total}\left(Y,\hat{Y}\right)}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z_{SS}^{t+1}} \cdot \frac{\partial Z_{SS}^{t+1}}{\partial {H^{GCN}}^{S\,t+1}} \cdot \frac{\partial {H^{GCN}}^{S\,t+1}}{\partial \omega_{gcn}^{t}}$ of the corresponding parameters of the FC layer can be obtained, and the parameters of the GCN subnetwork can be updated according to $\widetilde{\omega}_{gcn}^{t+1} = \omega_{gcn}^{t} - \eta_{gcn} \nabla \omega_{gcn}^{t+1}$. After updating all the subnetworks, the parameters of the FC layer are updated according to $\widetilde{\omega}_{F}^{t+1} = \omega_{F}^{t} - \eta_{F} \nabla \omega_{F}^{t+1}$ so that the FC layer can learn effective spatio-temporal–spectral fusion features.

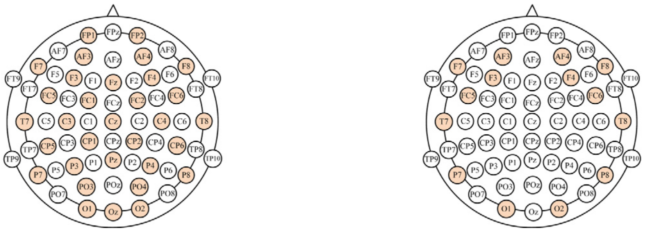
5. Experiments

5.1. Experimental datasets and setting

5.1.1. Datasets

Two widely used public emotion datasets, DEAP and DREAMER, were utilized to validate the effectiveness of the proposed MTLFuseNet model. The DEAP dataset contains 32 subjects, and each subject completed 40 experimental trials. Each trial required subjects to watch a 1-min music video to induce the corresponding emotional state. While watching music videos, the EEG signals and other peripheral physiological signals of the subjects were recorded. EEG signals were collected on 32 electrodes with a sampling rate of 512 Hz. After each trial, the subjects self-rated their emotional state from five dimensions: valence, arousal, dominance, liking, and familiarity. Among them, familiarity used a 5-point scale, and the rest used a 9-point scale. The DEAP dataset provides both raw data and preprocessed data. The preprocessed data are downsampled from 512 Hz to 128 Hz, and a bandpass frequency filter of 4.0–45.0 Hz was used to remove artifacts.

The DREAMER dataset contains 23 subjects, and each subject completed 18 experimental trials. Each trial was conducted by watching a video to induce the corresponding emotional state. The length of the 18 video clips varied, ranging from 65 s to 393 s. While watching the music video, the EEG and ECG signals of the subjects were recorded. The EEG signal acquisition device is a low-cost, portable and wearable wireless device that is collected on 14 electrodes with a sampling rate of 128 Hz. After each trial, the subjects needed to self-rate their emotional state from the dimensions of valence, arousal and dominance on a scale of 1–5. The DREAMER dataset provides preprocessed data, a bandpass frequency filter from 4.0 to 30.0 Hz was applied to remove artifacts.


FP1 FPz FP2
AF7 AF3 AFz AF4 AF8
F7 F5 F3 F1 Fz F2 F4 F6 F8
FT9 FT7 FC5 FC3 FC1 FCz FC2 FC4 FC6 FT8 FT10
T7 C5 C3 C1 Cz C2 C4 C6 T8
TP9 TP7 CP5 CP3 CP1 CPz CP2 CP4 CP6 TP8 TP10
P7 P5 P3 P1 Pz P2 P4 P6 P8
PO7 PO3 POz PO4 PO8
O1 Oz O2
FP1 FPz FP2
AF7 AF3 AFz AF4 AF8
F7 F5 F3 F1 Fz F2 F4 F6 F8
FT9 FT7 FC5 FC3 FC1 FCz FC2 FC4 FC6 FT8 FT10
T7 C5 C3 C1 Cz C2 C4 C6 T8
TP9 TP7 CP5 CP3 CP1 CPz CP2 CP4 CP6 TP8 TP10
P7 P5 P3 P1 Pz P2 P4 P6 P8
PO7 PO3 POz PO4 PO8
O1 Oz O2

(a) EEG electrode position visualization of DEAP

(b) EEG electrode position visualization of DREAMER

Fig. 5. EEG electrode position visualization.

Table 1
DEAP:MTLFuseNet experiment parameters.

Parameter	Value
Number of frequency bands	4
Spatio-temporal latent feature dimension	128
Spatio-spectral latent feature dimension	512
Number of channels	32
Dimension of Adjacency Matrix	32 × 32
Dropout	0.2
Learning Rate	1e-4

Table 2
DREAMER: MTLFuseNet experiment parameters.

Parameter	Value
Number of frequency bands	3
Spatio-temporal latent feature dimension	128
Spatio-spectral latent feature dimension	384
Number of channels	14
Dimension of Adjacency Matrix	14 × 14
Dropout	0.2
Learning Rate	1e-4

5.1.2. Parameter setting

There are 1280 samples in the DEAP dataset (32 subjects × 40 trials), including 32 EEG channels. The channel locations are shown in Fig. 5(a), and spatio-temporal coding is performed according to Fig. 2. Four frequency bands θ, α, β and γ were extracted in the construction of the spatio-spectral graph sequence in this study. The detailed parameter settings of the MTLFuseNet model for the DEAP dataset are shown in Table 1.

The DREAMER dataset contains 414 samples (23 subjects × 18 trials), with 14 EEG channels, as shown in Fig. 5(b), and spatio-temporal encoding according to Fig. 2. Since the EEG signals in the DREAMER dataset were filtered by a bandpass filter from 4.0–30.0 Hz, only θ, α, and β bands were extracted for the construction of the spatio-spectral graph sequence. The detailed parameter settings of MTLFuseNet for the DREAMER dataset are shown in Table 2.

MTLFuseNet consists of two subnets, the VAE network and GCN-GRU network, and performs three tasks based on the fusion of spatio-temporal–spectral latent features. The VAE encoder is composed of 4 layers of a convolutional neural network, the convolutional kernels of each layer are 3 × 3 × 128, 3 × 3 × 256, 3 × 3 × 256 and 3 × 3 × 512, and the stride is 1. The GCN-GRU network is composed of a GCN network and a GRU network. To find the optimal set of hyperparameters for three tasks, this study conducted an initial experiment of parameter search and obtained the following parameters β for the loss function $\pounds_{Total}$ defined in Eq. (22). The grid search algorithm performs on β_1, β_2, and β_3 in the set $\{0.1, 0.2, 0.3, 0.7\}$. The parameters β_1, β_2, β_3 are set to 0.7, 0.2 and 0.1.

5.1.3. Label processing

Two dimensions of arousal and valence were used to verify the proposed MTLFuseNet model in this study, as they can

Table 3
DEAP: benchmark comparison experiments on different evaluation metrics (mean ± std).

Model	Valence					Arousal				
	Accuracy	Recall	Precision	F1	ROC-AUC	Accuracy	Recall	Precision	F1	ROC-AUC
Wang et al. (2021) [55]	66.47 ± 8.75	–	–	–	–	69.79 ± 11.93	–	–	–	–
Priyasad et al. (2022) [53]	69.43 ± -	–	–	–	–	69.72 ± -	–	–	–	–
He et al. (2022) [56]	64.33 ± 7.06	–	–	–	–	63.25 ± 4.62	–	–	–	–
Zhang et al. (2023) [57]	64.66 ± 6.24	–	–	60.23 ± -	–	66.65 ± 8.16	–	–	59.57 ± -	–
EEGNet [58]*	63.67 ± 6.87	78.43 ± 19.15	63.44 ± 8.33	69.21 ± 11.36	58.67 ± 5.27	66.72 ± 9.21	81.07 ± 22.66	65.34 ± 12.66	70.91 ± 15.68	57.89 ± 5.75
FBCCNN [59]*	63.44 ± 7.09	84.91 ± 22.31	63.29 ± 10.22	70.30 ± 13.11	57.67 ± 6.35	65.55 ± 11.5	78.08 ± 29.14	66.28 ± 15.95	67.79 ± 21.06	56.02 ± 5.96
DGCNN [60]*	60.23 ± 8.6	66.34 ± 37.26	59.97 ± 19.28	57.87 ± 26.19	55.81 ± 7.18	64.38 ± 13.15	74.75 ± 38.12	61.15 ± 28.89	61.93 ± 30.35	54.73 ± 6.65
FBCNet [61]*	61.80 ± 8.31	80.68 ± 30.12	60.06 ± 19.23	65.96 ± 21.13	55.89 ± 7.33	65.55 ± 10.21	72.34 ± 36.19	59.89 ± 25.3	63.02 ± 28.61	55.05 ± 6.38
Tsception [62]*	66.72 ± 9.83	**86.40 ± 29.16**	63.55 ± 18.63	71.89 ± 22.11	52.99 ± 6.54	67.42 ± 9.81	**86.46 ± 27.43**	66.81 ± 16.15	72.97 ± 19.03	53.66 ± 6.87
STNet [63]*	64.84 ± 6.5	84.83 ± 16.15	65.15 ± 10.86	71.86 ± 11.32	60.42 ± 5.95	65.08 ± 9.99	82.34 ± 24.72	62.57 ± 16.35	69.88 ± 18.91	55.35 ± 5.75
MT-CNN [64]*	65.31 ± 7.37	81.77 ± 21.15	68.02 ± 13.39	70.99 ± 10.95	60.66 ± 8.22	67.34 ± 11.83	83.74 ± 25.53	67.90 ± 16.36	71.33 ± 18.36	57.80 ± 8.32
SSTD [65]*	68.28 ± 6.82	76.19 ± 20.06	**71.02 ± 11.7**	70.97 ± 10.57	64.20 ± 7.57	71.48 ± 7.18	73.54 ± 27.81	**74.54 ± 13.51**	69.21 ± 20.96	62.52 ± 9.36
MTLFuseNet	**71.33 ± 5.24**	85.81 ± 11.35	70.49 ± 9.23	**76.46 ± 6.46**	**68.62 ± 5.1**	**73.28 ± 7.74**	83.51 ± 16.17	71.30 ± 13.64	**76.22 ± 12.98**	**68.04 ± 6.04**

* indicates the experiment results obtained by our own implementation.
– indicates the experiment results are not reported on that dataset.

best describe emotional states and have been widely used by researchers. The arousal dimension represents the degree of excitement or inhibition of emotion, while the valence dimension represents the positive or negative level of emotion. Label processing maps each dimension to a binary classification. The median of the arousal and valence dimension labels in DEAP is 5, while DREAMER is 3. Scores above or equal to the median were mapped to a high arousal/valence class (positive, 1), and scores below the median were mapped to a low arousal/valence class (negative, -1).

5.1.4. Cross validation

There are subject-dependent and subject-independent classifications according to the method of partitioning the dataset into training and testing sets. Subject-dependent classifications are trained and tested on data from the same subject, while subject-independent classifications are trained on data from different subjects and then tested on new subjects not included in the training set. Since the pattern of cortical features of the same individual is stable and fits principles of statistical learning [66], subject-dependent classifications often have better performance [67]. However, in practical applications, established models are usually applied directly to new individuals [68], so subject-independent classifications are more realistic. Therefore, in this study, a subject-independent leave-one-subject-out cross validation (LOSOCV) strategy [69] was adopted to validate the MTLFuseNet model. For the DEAP and DREAMER datasets, 31 and 22 subjects were used for training, respectively, leaving 1 subject (not included in the training set) for testing, and cross-validation was performed on all subjects.

5.1.5. Evaluation metrics

In this study, a variety of evaluation metrics were used to describe the classification results to verify the validity and reliability of the MTLFuseNet model from different perspectives. The evaluation metrics used are accuracy, precision, recall, F1 score and ROC-AUC.

5.1.6. Experimental setup

All the experiments were conducted on a desktop with 3.2 GHz Intel Core CPU i7 and 64 GB memory.

5.2. Benchmark comparison experiments

To validate the effectiveness of the proposed MTLFuseNet model, it is compared with some baseline methods, including EEGNet [58], FBCCNN [59], DGCNN [60], FBCNet [61], Tsception [62], STNet [63], MT-CNN [64] and SSTD [65]. According to the code provided by these baseline methods, we adopted the same LOSOCV validation strategy as MTLFuseNet and verified them in our experiment. The specific model descriptions are as follows:

- EEGNet [58]: A compact convolutional neural network for the field of brain–computer interfaces. The network structure used in this experiment is as follows: the convolution kernel size of block 1 is 64, and the number of channels is 8. The convolution kernel size of block 2 is 16, and the number of channels is 16.
- FBCCNN [59]: A spectrum dependent convolutional neural network for EEG emotion recognition. In this experiment, PSD features were extracted from the $\theta, \alpha, \beta, \gamma$ bands of the DEAP dataset and the θ, α, β bands of the DREAMER dataset and fed into the model. The model is composed of 7 convolutional layers, the size of the convolutional kernel is 3 × 3, the step size is 1, and the channel numbers of the 7 convolutional layers are 12, 32, 64, 128, 258, 128, and 32.
- DGCNN [60]: A novel multichannel EEG emotion recognition model based on dynamic graph convolutional neural network. In this experiment, the DE features were extracted from the $\theta, \alpha, \beta, \gamma$ bands of the DEAP dataset and the θ, α, β bands of the DREAMER dataset and fed into the model. The model consists of 2 graph convolution layers.
- FBCNet [61]: A novel filter-bank convolutional neural network which employs multiview data representation and spatial filtering to extract discriminative spectral features. In this experiment, 4 bands $\theta, \alpha, \beta, \gamma$ of the DEAP dataset and 3 bands θ, α, β of the DREAMER dataset were used. The network of this experiment consists of 1 convolutional layer, 1 temporal variance layer and 1 fully connected layer.
- Tsception [62]: A network capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. In this experiment, the number of multiscale 1D temporal kernels in the dynamic temporal layer is set to 15. The number of multiscale 1D spatial kernels in the asymmetric spatial layer is set to 15. The number of hidden nodes in the first fully connected layer is set to 32.
- STNet [63]: A spatio-temporal network for EEG-based emotion recognition. The network of this experiment consists of 3 convolutional layers, 1 separable Conv2d layer and 1 inception Conv2d layer.
- MT-CNN [64]: A multi-task convolutional neural network for emotion recognition based on EEG brain maps. In this experiment, the model is composed of 4 convolutional layers with a step size of 1, and the kernel sizes of the convolutional layers are 5 × 5, 4 × 4, 4 × 4 and 1 × 1.
- SSTD [65]: A novel spatio-temporal demographic network for EEG-based emotion recognition. In this experiment, the model consists of a 2-layer GRU and SPDNet, in which SPDNet contains 2 BiMap layers, 1 ReEig layer and 1 LogEig layer.

In addition, we compared recent studies [53,55–57] on EEG emotion recognition that also employ the LOSOCV validation

Table 4
DREAMER: benchmark comparison experiments on different evaluation metrics (mean ± std).

Model	Valence					Arousal				
	Accuracy	Recall	Precision	F1	ROC-AUC	Accuracy	Recall	Precision	F1	ROC-AUC
Wang et al. (2021) [55]	67.99 ± 6.34	–	–	–	–	76.57 ± 14.04	–	–	–	–
Priyasad et al. (2022) [53]	64.98 ± -	–	–	–	–	63.71 ± -	–	–	–	–
He et al. (2022) [56]	66.56 ± 10.04	–	–	–	–	63.69 ± 6.57	–	–	–	–
Zhang et al. (2023) [57]	68.18 ± -	–	–	59.29 ± -	–	67.12 ± -	–	–	60.31 ± -	–
EEGNet [58]*	70.53 ± 8.28	77.37 ± 30.63	75.22 ± 12.55	70.90 ± 20.81	59.71 ± 7.47	74.15 ± 13.04	92.13 ± 23.96	76.62 ± 13.26	80.37 ± 18.7	53.36 ± 6.43
FBCCNN [59]*	64.49 ± 10.15	82.72 ± 37.47	58.38 ± 26.15	65.75 ± 29.54	50.60 ± 1.8	73.43 ± 13.18	**100 ± 0**	73.08 ± 13.56	83.77 ± 9.02	51.45 ± 4.18
DGCNN [60]*	69.57 ± 7.47	76.59 ± 32.31	71.40 ± 19.89	69.35 ± 22.95	57.56 ± 7.69	76.33 ± 12.45	95.56 ± 10.68	77.34 ± 13.48	84.43 ± 9.46	57.41 ± 11.78
FBCNet [61]*	65.22 ± 11.26	89.48 ± 12.4	61.99 ± 12.79	72.16 ± 11.38	51.89 ± 7.59	74.15 ± 14.27	98.09 ± 4.53	74.23 ± 14.15	83.81 ± 9.89	54.01 ± 7.81
Tsception [62]*	62.07 ± 14.08	91.54 ± 25.22	60.78 ± 17.77	71.57 ± 19.74	51.46 ± 5.44	75.36 ± 14.68	89.66 ± 23.38	74.84 ± 20.8	80.59 ± 20.32	58.07 ± 9.44
STNet [63]*	67.39 ± 10.77	79.51 ± 37.97	65.29 ± 24.53	66.43 ± 29.19	55.33 ± 5.75	72.95 ± 14.24	95.65 ± 20.85	70.69 ± 20.34	80.64 ± 19.7	51.09 ± 3.6
MT-CNN [64]*	62.56 ± 12.89	**95.34 ± 20.84**	59.90 ± 18.42	72.76 ± 18.77	51.01 ± 3.44	72.47 ± 13.96	98.81 ± 5.69	72.71 ± 13.8	83.03 ± 9.78	50.34 ± 1.62
SSTD [65]*	76.81 ± 6.63	87.70 ± 22.24	76.67 ± 11.28	79.34 ± 16.1	68.00 ± 9.76	81.64 ± 10.78	95.12 ± 8.32	81.21 ± 11.7	87.18 ± 8.93	69.69 ± 13.63
MTLFuseNet	**80.43 ± 8.01**	89.39 ± 15.56	**79.97 ± 11.68**	**82.97 ± 10.61**	**75.49 ± 7.59**	**83.33 ± 11.24**	95.87 ± 9.59	**82.39 ± 11.96**	**88.12 ± 9.21**	**74.73 ± 13.68**

* indicates the experiment results obtained by our own implementation.
– indicates the experiment results are not reported on that dataset.

Table 5
DEAP: Spatio-Temporal and Spatio-Spectral submodel comparison experiments on different evaluation metrics (mean ± std).

Model	Valence					Arousal				
	Accuracy	Recall	Precision	F1	ROC-AUC	Accuracy	Recall	Precision	F1	ROC-AUC
Spatio-Spectral	63.83 ± 8.01	81.34 ± 24.52	63.74 ± 16.49	68.91 ± 16.60	58.89 ± 7.37	67.27 ± 11.52	81.13 ± 33.44	65.17 ± 22.21	67.62 ± 25.85	56.82 ± 6.34
Spatio-Temporal	70.08 ± 4.64	79.34 ± 12.58	**71.71 ± 9.63**	74.23 ± 6.38	68.40 ± 4.81	70.86 ± 6.18	78.23 ± 16.86	**72.45 ± 11.70**	73.79 ± 11.49	66.36 ± 4.76
MTLFuseNet	**71.33 ± 5.24**	**85.81 ± 11.35**	70.49 ± 9.23	**76.46 ± 6.46**	**68.62 ± 5.1**	**73.28 ± 7.74**	**83.51 ± 16.17**	71.30 ± 13.64	**76.22 ± 12.98**	**68.04 ± 6.04**

strategy and are experimentally validated on both the DEAP and DREAMER datasets. Their results are quoted directly from the literature to ensure a convincing comparison with the proposed MTLFuseNet model.

Table 3 and Fig. 6 show the results of the benchmark comparison experiments on the DEAP dataset. Compared with all baseline models, the proposed MTLFuseNet model improved the accuracy in the valence and arousal dimensions by 3.05%–11.1% and 1.8%–10.03%, respectively. Table 4 and Fig. 7 show the results of the benchmark comparison experiments on the DREAMER dataset. Compared with all baseline models, the proposed MTLFuseNet model improves the accuracy by 3.62%–18.36% and 1.69%–19.64% in the valence and arousal dimensions, respectively.

The MTLFuseNet model fused the spatio-temporal–spectral features of EEG signals with three tasks to form a more complementary and discriminative feature representation and construct a more robust model for emotion recognition. The experimental results show that the proposed MTLFuseNet model has good performance on both the DEAP and DREAMER datasets. The MTLFuseNet model achieves the best accuracies on the vast majority of subjects and has the best mean accuracy. In addition, the MTLFuseNet model has the most stable performance compared to the baseline methods. It is noteworthy that, from Table 3, Tsception and SSTD models have higher recall and precision on the valence and arousal dimensions of the DEAP dataset than the MTLFuseNet model, but other evaluation indicators are lower than the MTLFuseNet model. From Table 4, it can be seen that the MT-CNN and FBCCNN models obtained higher recall on the valence and arousal dimensions of the DREAMER dataset, respectively, but other evaluation indicators are lower than the MTLFuseNet model. There are two reasons for this phenomenon. First, the DEAP and DREAMER datasets are imbalanced. Second, the prediction results of the benchmark comparison models tend to be positive, while the prediction results of the negative classification are not ideal, resulting in high recall and low accuracy. According to the results of the benchmark comparison experiments, our proposed MTLFuseNet model achieved the best results in terms of accuracy, F1 and ROC-AUC, further demonstrating that the stability and generalization ability of MTLFuseNet model are superior to other models.

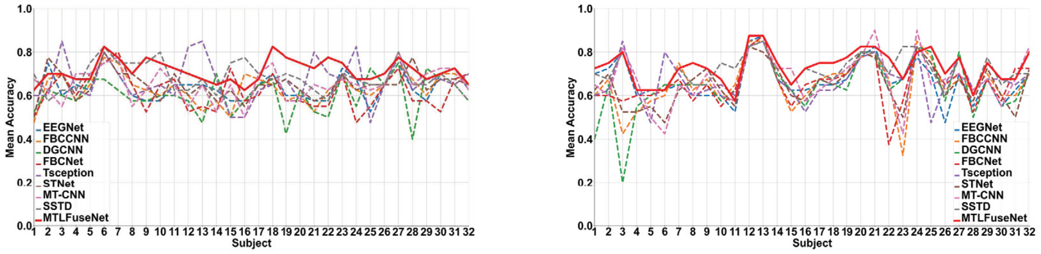
5.3. Spatio-temporal and spatio-spectral submodel comparison experiments

To verify the effects of each feature representation component of the MTLFuseNet model, spatio-temporal and spatio-spectral submodel comparison experiments were conducted. The models are described as follows:

- Spatio-Temporal submodel: VAE was employed to investigate the validity of spatio-temporal features. The VAE encoder is composed of 4 layers of a convolutional neural network, the convolutional kernels of each layer are 3 × 3 × 128, 3 × 3 × 256, 3 × 3 × 256 and 3 × 3 × 512, and the stride is 1. The network architecture of the decoder corresponds to that of the encoder. The loss function is defined as $\pounds_{Total}$ in Eq. (22), with the parameters β_1, β_2, β_3 being 0.7, 0.2 and 0.1, respectively.
- Spatio-Spectral submodel: GCN-GRU was employed to investigate the validity of the spatio-spectral features. The GCN-GRU network is composed of a GCN network and a GRU network. The loss function is $\pounds_{GCN-GRU} = \beta_1\pounds_{FL} + \beta_2\pounds_{TC}$, and the parameters β_1, β_2 are 0.7 and 0.3.

Table 5 and Fig. 8 show the results of the spatio-temporal and spatio-spectral submodel comparison experiments on the DEAP dataset. Comparing the experimental results of the two submodels, it can be found that the average accuracy of the spatio-temporal submodel is 6.25% and 3.59% higher than that of the spatio-spectral submodel in the valence and arousal dimensions, respectively. Comparing the MTLFuseNet model with the two submodels, it is found that the average accuracy of the proposed MTLFuseNet model is increased by 1.25%–7.5% and 2.42%–6.01% in the valence and arousal dimensions, respectively.

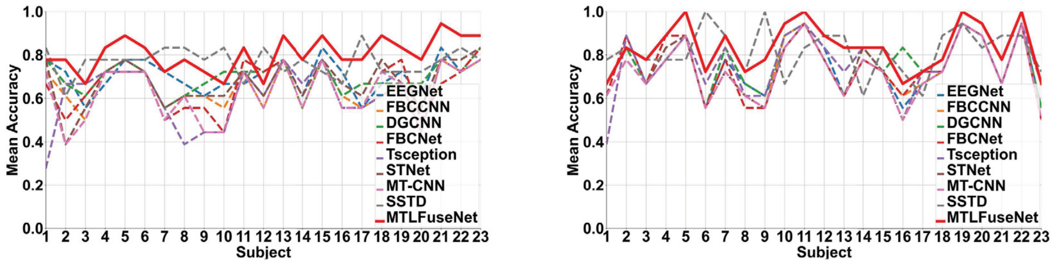
Table 6 and Fig. 9 show the results of the spatio-temporal and spatio-spectral submodel comparison experiments on the DREAMER dataset. Comparing the experimental results of the two submodels, it can be found that the average accuracy of the spatio-temporal submodel is 7.97% and 5.32% higher than that of the spatio-spectral submodel in the valence and arousal dimensions, respectively. Comparing the MTLFuseNet model with the two submodels, it is found that the average accuracy of the proposed MTLFuseNet model is increased by 1.93%–9.9% and 2.65%–7.97% in the valence and arousal dimensions, respectively.



(a) DEAP: average accuracy on each subject of benchmark comparison experiments on valence

(b) DEAP: average accuracy on each subject of benchmark comparison experiments on arousal

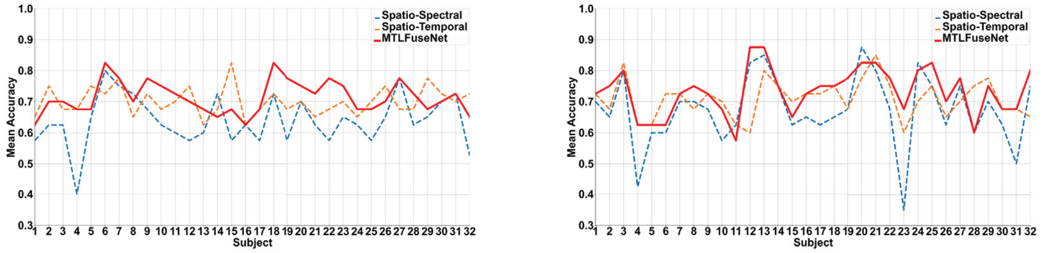
Fig. 6. DEAP: average accuracy on each subject of benchmark comparison experiments on valence and arousal.



(a) DREAMER: average accuracy on each subject of benchmark comparison experiments on valence

(b) DREAMER: average accuracy on each subject of benchmark comparison experiments on arousal

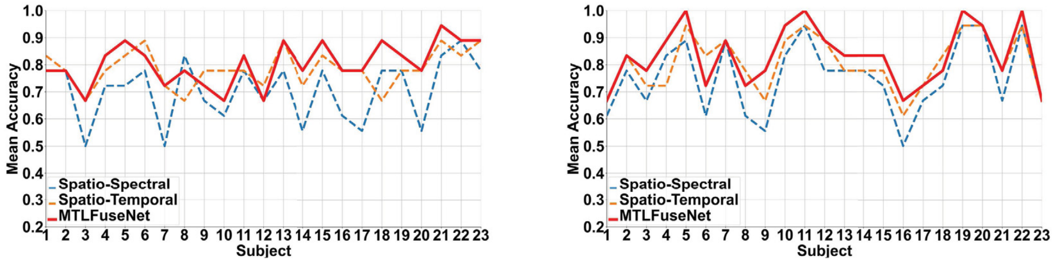
Fig. 7. DREAMER: average accuracy on each subject of benchmark comparison experiments on valence and arousal.



(a) DEAP: average accuracy on each subject of submodel experiments on valence

(b) DEAP: average accuracy on each subject of submodel experiments on arousal

Fig. 8. DEAP: average accuracy on each subject of Spatio-Temporal and Spatio-Spectral submodel comparison experiments on valence and arousal.



(a) DREAMER: average accuracy on each subject of submodel experiments on valence

(b) DREAMER: average accuracy on each subject of submodel experiments on arousal

Fig. 9. DREAMER: average accuracy on each subject of Spatio-Temporal and Spatio-Spectral submodel comparison experiments on valence and arousal.

Table 6
DREAMER: Spatio-Temporal and Spatio-Spectral submodel comparison experiments on different evaluation metrics (mean ± std).

Model	Valence					Arousal				
	Accuracy	Recall	Precision	F1	ROC-AUC	Accuracy	Recall	Precision	F1	ROC-AUC
Spatio-Spectral	70.53 ± 11.41	82.19 ± 28.65	69.92 ± 19.68	72.80 ± 21.66	60.19 ± 11.43	75.36 ± 13.38	93.11 ± 14.01	77.20 ± 14.11	83.05 ± 11	56.15 ± 11.07
Spatio-Temporal	78.50 ± 6.97	87.73 ± 10.8	78.27 ± 12.2	82.07 ± 8.7	75.41 ± 8.43	80.68 ± 10.44	91.90 ± 14.18	82.28 ± 11.37	85.81 ± 11.1	71.61 ± 11.62
MTLFuseNet	**80.43 ± 8.01**	**89.39 ± 15.56**	**79.97 ± 11.68**	**82.97 ± 10.61**	**75.49 ± 7.59**	**83.33 ± 11.24**	**95.87 ± 9.59**	**82.39 ± 11.96**	**88.12 ± 9.21**	**74.73 ± 13.68**

Table 7
DEAP: multi-task ablation experiments on different evaluation metrics (mean ± std).

Model	Valence					Arousal				
	Accuracy	Recall	Precision	F1	ROC_AUC	Accuracy	Recall	Precision	F1	ROC_AUC
MTLFuseNet-R	55.63 ± 9.76	49.06 ± 38.91	**71.55 ± 26.06**	46.06 ± 27.19	56.58 ± 5.32	54.92 ± 15.6	48.94 ± 39.2	73.20 ± 27.34	46.24 ± 28.28	58.09 ± 5.73
MTLFuseNet-C	57.11 ± 10.49	51.47 ± 38.07	62.64 ± 31.82	48.09 ± 29.02	57.69 ± 6.21	55.39 ± 15.53	49.56 ± 38.75	65.27 ± 31.66	46.61 ± 29.69	58.45 ± 6.29
MTLFuseNet-E	63.91 ± 6.35	77.43 ± 17.52	65.98 ± 10.19	69.54 ± 8.89	61.70 ± 4.48	64.77 ± 10.17	81.85 ± 17.11	65.48 ± 15.82	71.00 ± 13.86	60.21 ± 4.57
MTLFuseNet-RC	59.06 ± 8.47	59.34 ± 36.1	70.52 ± 20.03	54.98 ± 23.79	57.85 ± 5.61	60.08 ± 12.32	61.45 ± 35.78	**73.39 ± 22.79**	56.83 ± 25.25	59.23 ± 5.77
MTLFuseNet-RE	64.45 ± 7.4	78.31 ± 17.3	67.01 ± 11.06	70.26 ± 8.39	62.37 ± 5.66	66.95 ± 8.32	75.24 ± 24.13	71.24 ± 15.17	69.37 ± 15.34	62.32 ± 5.75
MTLFuseNet-CE	65.70 ± 6.61	74.73 ± 16.83	68.30 ± 11.29	69.75 ± 9.93	63.27 ± 5.87	67.73 ± 9.25	79.32 ± 21.17	67.96 ± 13.31	71.42 ± 14.82	61.67 ± 5.63
MTLFuseNet	**71.33 ± 5.24**	**85.81 ± 11.35**	70.49 ± 9.23	**76.46 ± 6.46**	**68.62 ± 5.1**	**73.28 ± 7.74**	**83.51 ± 16.17**	71.30 ± 13.64	**76.22 ± 12.98**	**68.04 ± 6.04**

The spatio-temporal and spatio-spectral submodel comparison experiments results show that the proposed MTLFuseNet model achieves good performance on both the DEAP and DREAMER datasets. The MTLFuseNet model obtains the best results on most subjects and the best mean accuracy. The spatio-temporal submodel performs better than the spatio-spectral submodel in the valence and arousal dimensions of the two datasets. This indicates that the effect of emotion classification based on the spatio-temporal latent features of unsupervised learning is better than that based on the spatio-spectral latent features of supervised learning. The classification accuracy of the MTLFuseNet model fused with spatio-temporal–spectral latent features is higher than that of submodels based on spatio-temporal features or spatio-spectral features. This indicates that the fusion of multidomain latent features can form more complementary and discriminative EEG features, which is beneficial for improving the emotion recognition effect of the model.

5.4. Multi-task ablation experiments

To evaluate the contribution of each basic task in our proposed model, experiments were conducted using the ablated MTLFuseNet model. The multi-task ablation experiments verified the effects of individual subtasks and multi-task combinations on the performance of EEG emotion recognition. Table 7, Fig. 10, Table 8 and Fig. 11 give the ablation experiment results on the DEAP and DREAMER datasets, respectively. MTLFuseNet-R, MTLFuseNet-C, and MTLFuseNet-E represent the ablated models that only consider the EEG data reconstruction task, latent feature vector construction task, and emotion recognition task, respectively. MTLFuseNet-RC, MTLFuseNet-RE, and MTLFuseNet-CE represent the ablated models that consider two tasks simultaneously. The two sub-tasks are assigned the same weight, with the weight coefficients being 0.5 and 0.5, respectively.

The results of the ablation experiments show that the emotion recognition task achieves the best performance in the single task, so it can be taken as the main task. In addition, the EEG data reconstruction task and the latent feature vector construction task also help the classification effect. Compared with the results of a single task, the combination of two tasks improves the accuracy of the model, indicating that these three tasks are related and can promote the model to learn more discriminative latent representations. In the case of two tasks, the performance of the MTLFuseNet-CE task on two datasets is the best, so it can be seen that the latent feature vector construction task and emotion recognition task are more helpful for the model classification effect, and higher weights can be assigned to these two tasks. The final results show that the MTLFuseNet model fused with all tasks achieves the best performance. This further proves the effectiveness of our proposed emotion recognition model based on deep latent feature fusion of EEG signals and multi-task learning.

5.5. Loss comparison experiments

To further verify the influence of different loss functions on the classification results, this study verified the effects of focal loss and triplet-center loss through experiments. The specific models are described as follows:

LVC-Loss: The network architecture and parameters are the same as in the MTLFuseNet model. The loss function is adopted as $\pounds_{LVF} = \beta_1 \pounds_{CE} + \beta_2 \pounds_{VAE}$, where $\pounds_{CE}$ is the cross entropy loss, and the parameters β_1, β_2 are 0.7 and 0.3, respectively.

LVF-Loss: The network architecture and parameters are the same as in the MTLFuseNet model. The loss function is adopted as $\pounds_{LVF} = \beta_1 \pounds_{FL} + \beta_2 \pounds_{VAE}$, where the parameters β_1, β_2 are 0.7 and 0.3.

Total-Loss: The MTLFuseNet model uses the $\pounds_{Total}$ loss function defined by Eq. (22), where the parameters $\beta_1, \beta_2, \beta_3$ are 0.7, 0.2 and 0.1, respectively.

Table 9 and Fig. 12 show the comparison results of different loss functions on the DEAP dataset. Comparing the experimental results of the LVC-Loss and LVF-Loss models, it can be found that the average accuracy of the LVF-Loss model improved by 2.5% and 1.48% compared with the LVC-Loss model in valence and arousal, respectively. Comparing the experimental results of the Total-Loss and LVF-Loss models, the Total-Loss model has 2.27% and 2.97% higher average accuracy in the valence and arousal dimensions, respectively, than the LVF-Loss model.

Table 10 and Fig. 13 show the results of the loss comparison experiments on the DREAMER dataset. Comparing the experimental results of the LVC-Loss and LVF-Loss models, it can be found that the average accuracy of the LVF-Loss model improved by 4.11% and 2.42% compared with the LVC-Loss model in valence and arousal, respectively. Comparing the experimental results of the Total-Loss and LVF-Loss models, the Total-Loss model has 2.17% and 3.86% higher accuracy in the valence and arousal dimensions, respectively, than the LVF-Loss model.

There is a sample imbalance problem in the DEAP and DREAMER datasets. According to the experimental results of the LVC-Loss and LVF-Loss models, LVF-Loss performs better than LVC-Loss in the valence and arousal dimensions of the two datasets. This indicates that the performance of the model has been improved by replacing the cross-entropy loss function with the focal loss function since the focal loss can solve the problem of unbalanced samples in the dataset. Comparing the experimental results of Total-Loss and LVF-Loss, it can be found that Total-Loss performs better than LVF-Loss in the valence and arousal dimensions of the two datasets. This indicates that adding the triplet-center loss function makes the fused EEG latent spatio-temporal–spectral features more separable, thus improving the classification performance of the model. The comparison of the loss functions shows that the $\pounds_{Total}$ loss function used in the MTLFuseNet model has achieved the best classification performance on the DEAP and DREAMER datasets.

5.6. Discussion

To evaluate the feasibility and pseudo-online performance of MTLFuseNet, this study performed sufficient benchmark comparison experiments, spatio-temporal and spatio-spectral submodel comparison experiments, multi-task ablation experiments and loss comparison experiments on two public datasets: DEAP and

Table 8
DREAMER: multi-task ablation experiments on different evaluation metrics (mean ± std).

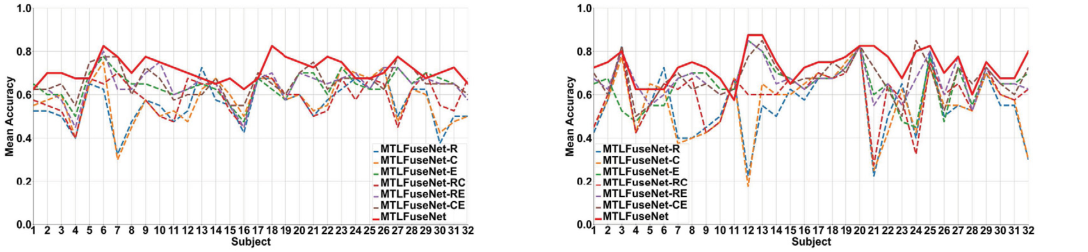
Model	Valence					Arousal				
	Accuracy	Recall	Precision	F1	ROC_AUC	Accuracy	Recall	Precision	F1	ROC_AUC
MTLFuseNet-R	64.73 ± 13.36	68.43 ± 34.51	78.76 ± 17.77	64.22 ± 22.99	62.20 ± 8.82	66.18 ± 17.96	67.09 ± 35.15	**87.00 ± 12.65**	67.38 ± 26.18	59.31 ± 9.57
MTLFuseNet-C	65.22 ± 14.33	68.23 ± 34.13	79.57 ± 17.49	64.66 ± 22.67	62.54 ± 9.5	67.87 ± 15.71	70.07 ± 32.06	85.91 ± 13.55	70.32 ± 23.11	61.48 ± 11.72
MTLFuseNet-E	74.40 ± 9.44	**91.19 ± 19.76**	74.75 ± 14.97	78.96 ± 13.15	68.36 ± 5.96	75.85 ± 12.15	90.29 ± 23.17	78.00 ± 11.68	81.12 ± 17.43	55.97 ± 6.91
MTLFuseNet-RC	67.87 ± 10.98	67.35 ± 32.83	78.59 ± 23.37	66.01 ± 22.64	66.08 ± 7.95	71.98 ± 16.54	73.39 ± 31.1	82.39 ± 21.32	73.93 ± 23.21	67.24 ± 13.03
MTLFuseNet-RE	75.60 ± 9.87	81.01 ± 23.66	**80.31 ± 14.57**	77.23 ± 14.04	72.16 ± 7.16	78.99 ± 10.72	88.79 ± 16.21	83.86 ± 13.08	84.37 ± 10.28	67.30 ± 10.82
MTLFuseNet-CE	77.05 ± 8.59	88.34 ± 16.62	77.62 ± 13.95	80.74 ± 10.63	73.09 ± 6.3	81.16 ± 11.45	94.29 ± 13.17	80.85 ± 11.89	86.39 ± 10.9	69.10 ± 14.28
MTLFuseNet	**80.43 ± 8.01**	89.39 ± 15.56	79.97 ± 11.68	**82.97 ± 10.61**	**75.49 ± 7.59**	**83.33 ± 11.24**	**95.87 ± 9.59**	82.39 ± 11.96	**88.12 ± 9.21**	**74.73 ± 13.68**

Table 9
DEAP: loss comparison experiments on different evaluation metrics (mean ± std).

Model	Valence					Arousal				
	Accuracy	Recall	Precision	F1	ROC-AUC	Accuracy	Recall	Precision	F1	ROC-AUC
LVC-Loss	66.56 ± 7.34	**88.62 ± 18.36**	65.77 ± 10.74	73.20 ± 12.62	62.33 ± 4.61	68.83 ± 9.29	82.61 ± 26.04	66.58 ± 18.55	71.38 ± 19.65	60.34 ± 6.64
LVF-Loss	69.06 ± 4.66	83.42 ± 12.06	68.68 ± 8.33	74.52 ± 6.51	66.47 ± 4.41	70.31 ± 9.06	83.01 ± 22.16	67.77 ± 18.14	73.27 ± 17.84	64.63 ± 5.85
Total-Loss	**71.33 ± 5.24**	85.81 ± 11.35	**70.49 ± 9.23**	**76.46 ± 6.46**	**68.62 ± 5.1**	**73.28 ± 7.74**	**83.51 ± 16.17**	**71.30 ± 13.64**	**76.22 ± 12.98**	**68.04 ± 6.04**

Table 10
DREAMER: loss comparison experiments on different evaluation metrics (mean ± std).

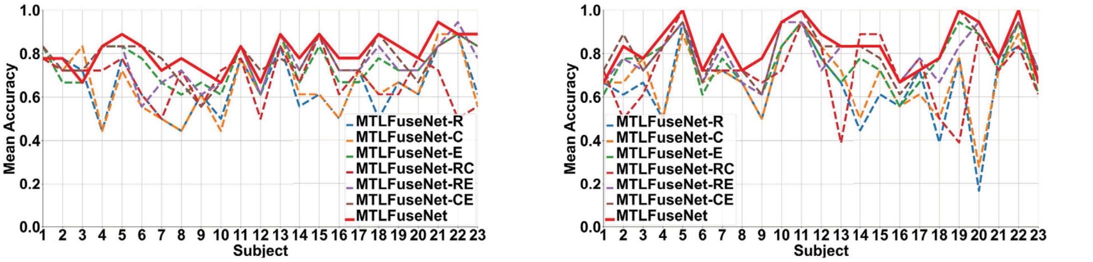
Model	Valence					Arousal				
	Accuracy	Recall	Precision	F1	ROC-AUC	Accuracy	Recall	Precision	F1	ROC-AUC
LVC-Loss	74.15 ± 8.64	82.75 ± 29.69	78.77 ± 13.46	74.93 ± 19.02	64.62 ± 7.39	77.05 ± 13.84	93.97 ± 14.19	77.31 ± 13.33	84.07 ± 12.2	60.01 ± 9.91
LVF-Loss	78.26 ± 8.19	84.49 ± 22.88	79.28 ± 10.57	79.39 ± 14.73	71.63 ± 7.79	79.47 ± 12.91	95.72 ± 7.54	79.54 ± 14.02	86.16 ± 9.48	69.31 ± 14.43
Total-Loss	**80.43 ± 8.01**	**89.39 ± 15.56**	**79.97 ± 11.68**	**82.97 ± 10.61**	**75.49 ± 7.59**	**83.33 ± 11.24**	**95.87 ± 9.59**	**82.39 ± 11.96**	**88.12 ± 9.21**	**74.73 ± 13.68**



(a) DEAP: average accuracy on each subject of multi-task ablation experiments on valence

(b) DEAP: average accuracy on each subject of multi-task ablation experiments on arousal

Fig. 10. DEAP: average accuracy on each subject of multi-task ablation comparison experiments on valence and arousal.

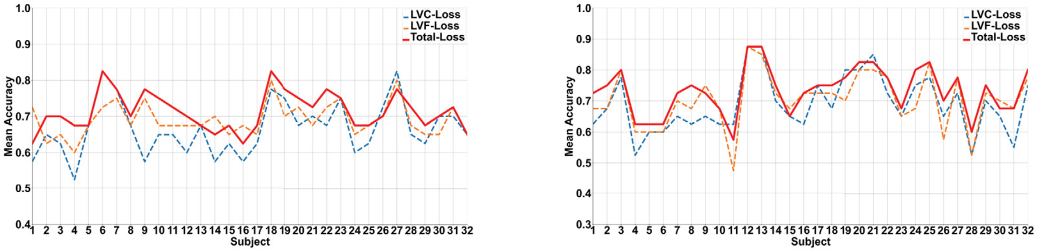


(a) DREAMER: average accuracy on each subject of multi-task ablation experiments on valence

(b) DREAMER: average accuracy on each subject of multi-task ablation experiments on arousal

Fig. 11. DREAMER: average accuracy on each subject of multi-task ablation comparison experiments on valence and arousal.

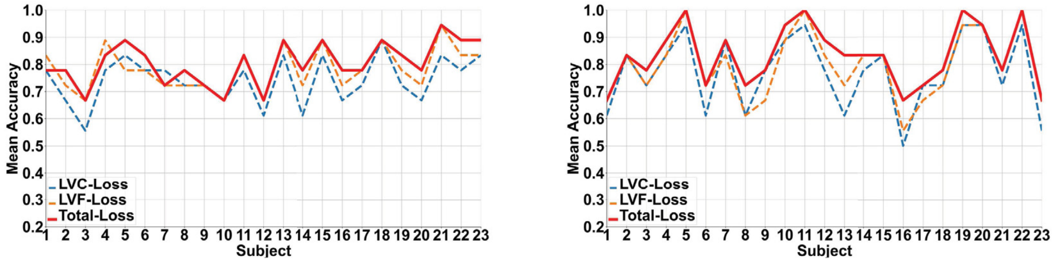
DREAMER. Compared with baseline models and SOTA methods, the MTLFuseNet model has the best emotion recognition effect. The outstanding classification accuracy of MTLFuseNet is first due to the ability of MTLFuseNet to extract more complementary and discriminative EEG emotion features in both unsupervised and supervised ways. From spatio-temporal and spatio-spectral submodel comparison experiments, each component of the MTLFuseNet model is validated. The experimental results show that emotion classification based on unsupervised learned spatio-temporal latent features performed better than that based on supervised learned spatio-spectral features. The MTLFuseNet model fused with spatio-temporal–spectral latent features has the highest classification accuracy, indicating



(a) DEAP: average accuracy on each subject of loss comparison experiments on valence

(b) DEAP: average accuracy on each subject of loss comparison experiments on arousal

Fig. 12. DEAP: average accuracy on each subject of loss comparison experiments on valence and arousal.



(a) DREAMER: average accuracy on each subject of loss comparison experiments on valence

(b) DREAMER: average accuracy on each subject of loss comparison experiments on arousal

Fig. 13. DREAMER: average accuracy on each subject of loss comparison experiments on valence and arousal.

Table 11
Time complexity of training (Ttrain) and prediction (Tpred) for all methods in seconds.

Model	DEAP		DREAMER	
	Ttrain	Tpred	Ttrain	Tpred
Spatio-Temporal	4.0053	1.4321	3.8315	0.6345
Spatio-Spectral	0.0765	0.0630	0.0631	0.0121
EEGNET	1.5581	0.4596	0.6651	0.1057
FBCCNN	0.0987	0.0299	0.0985	0.0153
DGCNN	0.0024	0.0008	0.0019	0.0006
FBCNET	0.6371	0.2753	0.2837	0.0428
Tsception	2.8130	1.1588	2.6042	0.5128
STNet	2.5554	0.8988	2.5180	0.4585
MT-CNN	0.0553	0.0171	0.0433	0.013
SSTD	0.6049	0.4741	0.1043	0.0822
MTLFuseNet	4.1389	1.5316	3.9379	0.6552

that the fusion of spatial, temporal and spectral multidomain latent features can form more complementary and discriminative EEG features, which is conducive to improving the effect of the emotion recognition model.

The model performance is also improved due to multi-task learning. The MTLFuseNet model has three emotion-related tasks: emotion recognition, metric learning-based latent feature vector construction and EEG data reconstruction. MTLFuseNet trains these three related tasks together, making different tasks complement and promote each other to improve the accuracy and generalization ability of the model. Through the multi-task ablation experiment, it is not difficult to find that the performance of the emotion recognition task is the best on a single task. The combination of two tasks improves the classification performance of the single task. Finally, the MTLFuseNet model combining the three tasks achieves the best classification performance, further proving the validity of our proposed multi-task emotion recognition model. Moreover, the use of focal loss and triplet-center loss is also beneficial. Loss comparison experiments show that the performance of the model is improved when the focal loss function is used to replace the cross entropy loss function. This indicates that the focal loss can improve the problem of poor emotion recognition effect caused by sample class imbalance. The triplet-center loss can make the fused spatio-temporal–spectral latent features more separable, thereby improving the accuracy of the emotion recognition model.

In addition, we calculated the training time and testing time of MTLFuseNet with spatio-temporal and spatio-spectral submodes and eight benchmark comparison methods, as shown in Table 11. According to Table 11, although MTLFuseNet requires a long training time, the prediction times of all testing trials on the DEAP and DREAMER datasets are 1.5316 s and 0.6552 s, respectively, which meet the time requirements of online applications. The above experimental results show that the proposed MTLFuseNet model is suitable for EEG-based emotion recognition.

6. Conclusions

In this study, a novel emotion recognition model based on deep latent spatio-temporal–spectral feature fusion of EEG signals and multi-task learning was proposed. To better obtain the spatio-temporal information of EEG signals, the original EEG signals were encoded into a spatio-temporal data representation, and then the spatio-temporal encoded data were sent to the VAE network to learn the spatio-temporal latent features of EEG signals in an unsupervised manner. At the same time, the spatial correlation information between EEG channels in different frequency bands was extracted by GCN, and then the spatial features extracted in different frequency bands were input into the GRU network to obtain the spatio-spectral features of EEG signals. After that, the spatio-temporal and spatio-spectral latent features extracted by the two subnetworks were fused to form the spatio-temporal–spectral features of EEG signals for emotion recognition. The proposed multi-task model was trained for three emotion-related tasks: emotion recognition task, metric learning-based latent feature vector construction task and EEG data reconstruction task. To make the latent features more separable and consider the problem of sample class imbalance in emotion datasets, this study introduced the triplet-center loss and focal loss. Based on three tasks, the total loss $\pounds_{Total}$, which consists of VAE loss, triplet-center loss and focal loss, was defined as the loss function of multi-task learning, and then an end-to-end MTLFuseNet model for EEG emotion recognition was established. Finally, this study validated the MTLFuseNet model on the DEAP and DREAMER datasets using the LOSOCV strategy. Through sufficient benchmark comparison experiments, spatio-temporal and spatio-spectral submodel comparison experiments, multi-task ablation experiments and loss comparison experiments, the experimental results show that the MTLFuseNet model proposed in this study has the best classification performance.

CRediT authorship contribution statement

Rui Li: Writing – original draft, Validation, Conceptualization. **Chao Ren:** Writing – review & editing, Software, Conceptualization. **Yiqing Ge:** Writing – review & editing. **Qiqi Zhao:** Writing – review & editing. **Yikun Yang:** Writing – review & editing. **Yuhan Shi:** Writing – review & editing. **Xiaowei Zhang:** Writing – review & editing. **Bin Hu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFA0706200), in part by the National Natural Science Foundation of China (Grant No. 62072219, No. 62227807), in part by the Natural Science Foundation of Gansu Province, China (Grant No. 22JR5RA401), and in part by the Fundamental Research Funds for the Central Universities, China (No. lzujbky-2022-ey13).

References

[1] Rodolfo Pavez, Jaime Diaz, Jeferson Arango-Lopez, Danay Ahumada, Carolina Mendez-Sandoval, Fernando Moreira, Emo-mirror: a proposal to support emotion recognition in children with autism spectrum disorders, Neural Comput. Appl. (2021) 1–12.

[2] Jose Maria Garcia-Garcia, Victor MR Penichet, Maria D Lozano, Anil Fernando, Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions, Univ. Access Inform. Soc. 21 (4) (2022) 809–825.

[3] Utkarsh Tripathi, Vinay Chamola, Alireza Jolfaei, Ananthakrishna Chintanpalli, et al., Advancing remote healthcare using humanoid and affective systems, IEEE Sens. J. 22 (18) (2021) 17606–17614.

[4] Jillian M Murphy, Joanne M Bennett, Xochitl de la Piedad Garcia, Megan L Willis, Emotion recognition and traumatic brain injury: A systematic review and meta-analysis, Neuropsychol. Review 32 (3) (2022) 520–536.

[5] Jesús Joel Rivas, Maria del Carmen Lara, Luis Castrejon, Jorge Hernandez-Franco, Felipe Orihuela-Espina, Lorena Palafox, Amanda Williams, Nadia Bianchi-Berthouze, Luis Enrique Sucar, Multi-label and multimodal classifier for affective states recognition in virtual rehabilitation, IEEE Trans. Affect. Comput. 13 (3) (2022) 1183–1194.

[6] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, Maja Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (1) (2011) 42–55.

[7] Saif Hassani, Ibrahim Bafadel, Abdelrahman Bekhatro, Ebraheim Al Blooshi, Soha Ahmed, Mahmoud Alahmad, Physiological signal-based emotion recognition system, in: 2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences, ICETAS, IEEE, 2017, pp. 1–5.

[8] Xiaowei Zhang, Jinyong Liu, Jian Shen, Shaojie Li, Kechen Hou, Bin Hu, Jin Gao, Tong Zhang, Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine, IEEE Trans. Cybern. 51 (9) (2020) 4386–4399.

[9] Soraia M. Alarcao, Manuel J. Fonseca, Emotions recognition using EEG signals: A survey, IEEE Trans. Affect. Comput. 10 (3) (2019) 374–393.

[10] Robert Jenke, Angelika Peer, Martin Buss, Feature extraction and selection for emotion recognition from EEG, IEEE Trans. Affect. Comput. 5 (3) (2014) 327–339.

[11] Xiang Li, Dawei Song, Peng Zhang, Yazhou Zhang, Yuexian Hou, Bin Hu, Exploring EEG features in cross-subject emotion recognition, Front. Neurosci. 12 (2018) 162.

[12] Zeynab Mohammadi, Javad Frounchi, Mahmood Amiri, Wavelet-based emotion recognition system using EEG signal, Neural Comput. Appl. 28 (8) (2017) 1985–1990.

[13] Fabian Parsia George, Istiaque Mannafee Shaikat, Prommy Sultana Ferdawoos, Mohammad Zavid Parvez, Jia Uddin, Recognition of emotional states using EEG signals based on time-frequency analysis and SVM classifier, Int. J. Electr. Comput. Eng. (2088-8708) 9 (2) (2019).

[14] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, Jocelyn Faubert, Deep learning-based electroencephalography analysis: a systematic review, J. Neural Eng. 16 (5) (2019) 051001.

[15] Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, Pekka Marttinen, EEG based emotion recognition: A tutorial and review, ACM Comput. Surv. 55 (4) (2022) 79.

[16] Michael X. Cohen, Where does EEG come from and what does it mean? Trends Neurosci. 40 (4) (2017) 208–218.

[17] Diederik P. Kingma, Max Welling, Auto-encoding variational bayes, 2013, arXiv preprint arXiv:1312.6114.

[18] Ahmed M. Abdelhameed, Magdy Bayoumi, Semi-supervised EEG signals classification system for epileptic seizure detection, IEEE Signal Process. Lett. 26 (12) (2019) 1922–1926.

[19] Thomas N. Kipf, Max Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.

[20] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.

[21] Rui Xia, Yang Liu, A multi-task learning framework for emotion recognition using 2D continuous space, IEEE Trans. Affect. Comput. 8 (1) (2017) 3–14.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[23] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, Xiang Bai, Triplet-center loss for multi-view 3D object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1945–1954.

[24] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, Ioannis Patras, DEAP: A database for emotion analysis using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2011) 18–31.

[25] Stamos Katsigiannis, Naeem Ramzan, DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices, IEEE J. Biomed. Health Inf. 22 (1) (2017) 98–107.

[26] Fu Yang, Xingcong Zhao, Wenge Jiang, Pengfei Gao, Guangyuan Liu, Multi-method fusion of cross-subject emotion recognition based on high-dimensional EEG features, Front. Comput. Neurosci. 13 (2019) 53.

[27] Md Mustafizur Rahman, Ajay Krishno Sarkar, Md Amzad Hossain, Mohammad Ali Moni, EEG-based emotion analysis using non-linear features and ensemble learning approaches, Expert Syst. Appl. 207 (2022) 118025.

[28] Nattapong Thammasan, Ken-ichi Fukui, Masayuki Numao, Application of deep belief networks in eeg-based dynamic music-emotion recognition, in: 2016 International Joint Conference on Neural Networks, IJCNN, IEEE, 2016, pp. 881–888.

[29] Anubhav, Debarshi Nath, Mrigank Singh, Divyashikha Sethia, Diksha Kalra, S. Indu, An efficient approach to eeg-based emotion recognition using lstm network, in: 2020 16th IEEE International Colloquium on Signal Processing & Its Applications, CSPA, IEEE, 2020, pp. 88–92.

[30] Zhe Wang, Tianhao Gu, Yiwen Zhu, Dongdong Li, Hai Yang, Wenli Du, FLDNet: Frame-level distilling neural network for EEG emotion recognition, IEEE J. Biomed. Health Inf. 25 (7) (2021) 2533–2544.

[31] Dongmin Huang, Sentao Chen, Cheng Liu, Lin Zheng, Zhihang Tian, Dazhi Jiang, Differences first in asymmetric brain: A bi-hemisphere discrepancy convolutional neural network for EEG emotion recognition, Neurocomputing 448 (2021) 140–151.

[32] Wenhui Guo, Guixun Xu, Yanjiang Wang, Horizontal and vertical features fusion network based on different brain regions for emotion recognition, Knowl.-Based Syst. 247 (2022) 108819.

[33] Shuaiqi Liu, Zeyao Wang, Yanling An, Jie Zhao, Yingying Zhao, Yu-Dong Zhang, EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network, Knowl.-Based Syst. 265 (2023) 110372.

[34] Kristen A. Lindquist, Lisa Feldman Barrett, A functional architecture of the human brain: emerging insights from the science of emotion, Trends in Cognitive Sciences 16 (11) (2012) 533–540.

[35] Wei-Long Zheng, Bao-Liang Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, IEEE Trans. Auton. Ment. Dev. 7 (3) (2015) 162–175.

[36] Peiyang Li, Huan Liu, Yajing Si, Cunbo Li, Fali Li, Xuyang Zhu, Xiaoye Huang, Ying Zeng, Dezhong Yao, Yangsong Zhang, et al., EEG based emotion recognition by combining functional connectivity network and local activations, IEEE Trans. Biomed. Eng. 66 (10) (2019) 2869–2881.

[37] Salma Alhagry, Aly Aly Fahmy, Reda A. El-Khoribi, Emotion recognition based on EEG using LSTM recurrent neural network, Int. J. Adv. Comput. Sci. Appl. 8 (10) (2017) 355–358.

[38] Zhe Wang, Yongxiong Wang, Chuanfei Hu, Zhong Yin, Yu Song, Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model, IEEE Sens. J. 22 (5) (2022) 4359–4368.

[39] Fatih Demir, Nebras Sobahi, Siuly Siuly, Abdulkadir Sengur, Exploring deep learning features for automatic classification of human emotion using EEG rhythms, IEEE Sens. J. 21 (13) (2021) 14923–14930.

[40] Shuaiqi Liu, Xu Wang, Ling Zhao, Bing Li, Weiming Hu, Jie Yu, Yu-Dong Zhang, 3DCANN: a spatio-temporal convolution attention neural network for EEG emotion recognition, IEEE J. Biomed. Health Inf. 26 (11) (2021) 5321–5331.

[41] Minmin Miao, Longxin Zheng, Baoguo Xu, Zhong Yang, Wenjun Hu, A multiple frequency bands parallel spatial–temporal 3D deep residual learning framework for EEG-based emotion recognition, Biomed. Signal Process. Control 79 (2023) 104141.

[42] Lin Feng, Cheng Cheng, Mingyan Zhao, Huiyuan Deng, Yong Zhang, EEG-based emotion recognition using spatial-temporal graph convolutional LSTM with attention mechanism, IEEE J. Biomed. Health Inf. 26 (11) (2022) 5406–5417.

[43] Qunli Yao, Heng Gu, Shaodi Wang, Xiaoli Li, A feature-fused convolutional neural network for emotion recognition from multichannel EEG signals, IEEE Sens. J. 22 (12) (2022) 11954–11964.

[44] Dongdong Li, Li Xie, Bing Chai, Zhe Wang, Hai Yang, Spatial-frequency convolutional self-attention network for EEG emotion recognition, Appl. Soft Comput. 122 (2022) 108740.

[45] Horace B. Barlow, Unsupervised learning, Neural Comput. 1 (3) (1989) 295–311.

[46] Zhen Liang, Shigeyuki Oba, Shin Ishii, An unsupervised EEG decoding system for human emotion recognition, Neural Netw. 116 (2019) 257–268.

[47] Rich Caruana, Multitask learning, Mach. Learn. 28 (1997) 41–75.

[48] Mengshi Ge, Rui Mao, Erik Cambria, Explainable metaphor identification inspired by conceptual metaphor theory, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 10, 2022, pp. 10681–10689.

[49] Rui Mao, Xiao Li, Mengshi Ge, Erik Cambria, MetaPro: A computational metaphor processing model for text pre-processing, Inf. Fusion 86–87 (2022) 30–43.

[50] Yu Zhang, Qiang Yang, An overview of multi-task learning, Natl. Sci. Rev. 5 (1) (2018) 30–43.

[51] Phairot Autthasan, Rattanaphon Chaisaen, Thapanun Sudhawiyangkul, Phurin Rangpong, Suktipol Kiatthaveephong, Nat Dilokthanakul, Gun Bhakdisongkhram, Huy Phan, Cuntai Guan, Theerawit Wilaiprasitporn, MIN2net: End-to-end multi-task learning for subject-independent motor imagery EEG classification, IEEE Trans. Biomed. Eng. 69 (6) (2022) 2105–2118.

[52] Chang Li, Bin Wang, Silin Zhang, Yu Liu, Rencheng Song, Juan Cheng, Xun Chen, Emotion recognition from EEG based on multi-task learning with capsule network and attention mechanism, Comput. Biol. Med. 143 (2022) 105303.

[53] Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, Clinton Fookes, Affect recognition from scalp-EEG using channel-wise encoder networks coupled with geometric deep learning and multi-channel feature fusion, Knowl.-Based Syst. 250 (2022) 109038.

[54] Yang Li, Ji Chen, Fu Li, Boxun Fu, Hao Wu, Youshuo Ji, Yijin Zhou, Yi Niu, Guangming Shi, Wenming Zheng, GMSS: Graph-based multi-task self-supervised learning for EEG emotion recognition, IEEE Trans. Affect. Comput. (2022) http://dx.doi.org/10.1109/TAFFC.2022.3170428.

[55] Yixin Wang, Shuang Qiu, Xuelin Ma, Huiguang He, A prototype-based SPD matrix network for domain adaptation EEG emotion recognition, Pattern Recognit. 110 (2021) 107626.

[56] Zhipeng He, Yongshi Zhong, Jiahui Pan, An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition, Comput. Biol. Med. 141 (2022) 105048.

[57] Guanhua Zhang, Minjing Yu, Yongjin Liu, Guozhen Zhao, Dan Zhang, Wenming Zheng, SparseDGCNN: recognizing emotion from multichannel EEG signals, IEEE Trans. Affect. Comput. 14 (1) (2023) 537–548.

[58] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, Brent J Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013.

[59] Bo Pan, Wei Zheng, Emotion recognition based on EEG using generative adversarial nets and convolutional neural network, Comput. Math. Methods Med. 2021 (2021) 2520394.

[60] Tengfei Song, Wenming Zheng, Peng Song, Zhen Cui, EEG emotion recognition using dynamical graph convolutional neural networks, IEEE Trans. Affect. Comput. 11 (3) (2018) 532–541.

[61] Ravikiran Mane, Effie Chew, Karen Chua, Kai Keng Ang, Neethu Robinson, A Prasad Vinod, Seong-Whan Lee, Cuntai Guan, FBCNet: A multi-view convolutional neural network for brain-computer interface, 2021, arXiv preprint arXiv:2104.01233.

[62] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, Cuntai Guan, Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition, IEEE Trans. Affect. Comput. (2022) http://dx.doi.org/10.1109/TAFFC.2022.3169001.

[63] Zhi Zhang, Sheng-hua Zhong, Yan Liu, GANSER: A self-supervised data augmentation framework for EEG-based emotion recognition, IEEE Trans. Affect. Comput. (2022) http://dx.doi.org/10.1109/TAFFC.2022.3170369.

[64] Evgenii Rudakov, Lou Laurent, Valentin Cousin, Ahmed Roshdi, Régis Fournier, Amine Nait-ali, Taha Beyrouthy, Samer Al Kork, Multi-task CNN model for emotion recognition from EEG brain maps, in: 2021 4th International Conference on Bio-Engineering for Smart Technologies, BioSMART, IEEE, 2021, pp. 1–4.

[65] Rui Li, Chao Ren, Chen Li, Nan Zhao, Dawei Lu, Xiaowei Zhang, SSTD: A novel spatio-temporal demographic network for EEG-based emotion recognition, IEEE Trans. Comput. Soc. Syst. 10 (1) (2023) 376–387.

[66] Zhong Yin, Lei Liu, Jianing Chen, Boxi Zhao, Yongxiong Wang, Locally robust EEG feature selection for individual-independent emotion recognition, Expert Syst. Appl. 162 (2020) 113768.

[67] Yang Li, Wenming Zheng, Lei Wang, Yuan Zong, Zhen Cui, From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition, IEEE Trans. Affect. Comput. 13 (2) (2022) 568–578.

[68] Swapnil Bhosale, Rupayan Chakraborty, Sunil Kumar Kopparapu, Calibration free meta learning based approach for subject independent EEG emotion recognition, Biomed. Signal Process. Control 72 (2022) 103289.

[69] Steffen Walter, Jonghwa Kim, David Hrabal, Stephen Clive Crawcour, Henrik Kessler, Harald C Traue, Transsituational individual-specific biopsychological classification of emotions, IEEE Trans. Syst. Man Cybern. Syst. 43 (4) (2013) 988–995.