Multi-to-Single: Reducing Multimodal Dependency in Emotion Recognition Through Contrastive Learning

Yan-Kai Liu, Jinyu Cai, Bao-Liang Lu, Wei-Long Zheng*

Shanghai Jiao Tong University
{liu-yankai, cai_jinyu, bllu, weilong}@sjtu.edu.cn

Abstract

Multimodal emotion recognition is a crucial research area in the field of affective brain-computer interfaces. However, in practical applications, it is often challenging to obtain all modalities simultaneously. To deal with this problem, researchers focus on using cross-modal methods to learn multimodal representations with fewer modalities. However, due to the significant differences in the distribution of different modalities, it is challenging to enable any modality to fully learn multimodal features. To address this limitation, we propose a Multi-to-Single (M2S) emotion recognition model, leveraging contrastive learning and incorporating two innovative modules: 1) a spatial and temporal-sparse (STS) attention mechanism that enhances the encoders' ability to extract features from data; 2) a novel Multi-to-Multi Contrastive Predictive Coding (M2M CPC) that learns and fuses features across different modalities. In the final testing, we only use a single modality for emotion recognition, reducing the dependence on multimodal data. Extensive experiments on five public multimodal emotion datasets demonstrate that our model achieves the state-of-the-art performance in the cross-modal tasks and maintains multimodal performance using only a single modality.

Code — https://github.com/Arcee-LYK/Multi-to-Single.

Introduction

Emotion recognition is an important research direction in affective brain-computer interfaces (BCI) (Li et al. 2022). Since emotion induction experiments usually only require participants to sit still and watch emotion induction videos, compared to other brain-computer interface studies, emotion induction experiments are more likely to obtain more physiological signals, such as electroencephalography (EEG), eye movements (EYE), electrocardiogram (ECG), and peripheral physiological signals (PPS). This advantage greatly promotes the applications of multimodal models in the field of emotion recognition. By fusing features from multiple paired modalities, the performance of these models is significantly enhanced compared to unimodal models.

However, these methods have significant limitations in practical applications. Not all modalities can be obtained in practical scenarios limited by the experimental environment. It is especially common for EEG signals, which are greatly sensitive to the environment. Any external interference and tiny movements of the subjects can bring significant noise to the EEG signal, leading to a decrease in signal quality or even making the data unusable. As for the EYE signals, although their collection process is convenient compared to EEG signals, there are still some issues. Acquiring eye movement data is sensitive to distance, the eye tracker must remain within a certain proximity to the subject, and the subject's eyes must stay focused on the screen. This requirement can be challenging for subjects during extended signal collection periods. As for the ECG and PPS, they have similar issues to EEG signals as they are all collected using similar electrodes. Therefore, in practical situations, it is hard to obtain high-quality data from multiple modalities at the same time.

Cross-modal methods are proposed to solve this problem. Cross-modal learning allows models to use different modalities during training and testing. There have been many studies in computer vision and natural language processing (Zhang et al. 2023; Lan et al. 2023). However, in the field of affective BCI, the research is still in its early stages, and only a limited number of studies have explored the transfer of EEG to other modalities. These studies cannot achieve the conversion between any modality, nor do they fully utilize the multimodal features in the training set.

To achieve multimodal effects using unimodal data, we propose a Multi-to-Single (M2S) emotion recognition model. The model only requires paired, unlabeled multimodal data during the pre-training phase, we can fine-tune the model using unimodal data on downstream tasks. During pre-training, we extract features with two encoders for each modality, emotion-related (ER) and emotion-independent (EI) encoders, and minimize their mutual information. For EEG, ECG, and PPS data that use electrodes to collect, we propose a novel spatial and temporal-sparse (STS) attention mechanism to fully utilize the continuous temporal and spatial features of data. Contrastive learning methods are applied to fuse and align features from different modalities, including an improved version of Multi-to-Multi Contrastive Predictive Coding (M2M CPC) and InfoNCE. In the fine-tuning phase, we connect a classifier after the ER encoder for emotion recognition. We freeze the encoder and

*Corresponding author.

only tune the classifier using one modality. To verify the robustness of the model, we conduct experiments on five public multimodal emotion datasets, including cross-modal and unimodal experimental settings. We also compare our work with multimodal methods. The experimental results indicate that our model achieves state-of-the-art performance on multiple tasks.

The main contributions of this paper are as follows:

- We propose a novel M2S model for cross-modal emotion recognition, which can reduce dependence on multimodal data while achieving multimodal performance with a single modality.
- We introduce two novel modules: the STS attention mechanism and M2M CPC. The former can fully utilize EEG data, while the latter can learn and fuse features across different modalities, helping existing methods perform better.
- We systematically evaluate the performance of our model on multiple datasets and modalities, including EEG-EYE, EEG-ECG, and EEG-PPS. Numerous experimental results have demonstrated that our model has excellent robustness.

Related Work

Emotion Recognition

Emotion recognition is a common and important application in brain-computer interfaces. Researchers use various physiological signals and neural networks to determine the emotional state of subjects. The most common signal is the EEG signal. Compared to functional Magnetic Resonance Imaging (fMRI) and some embedded brain signal acquisition methods, the EEG acquisition process is safer, more convenient, and more efficient. Duan et. al. (2013) recognized emotions by extracting the differential entropy (DE) feature from EEG signals, which has five frequency bands (δ: 1-3 Hz, θ: 4-7 Hz, α: 8-13 Hz, β: 14-30 Hz, γ: 31-50 Hz). It has been widely proven to be the most effective feature in emotion recognition (Zheng and Lu 2015).

With the development of multimodal technology, the data needed for emotion recognition is not limited to a single modality. Zheng et. al. (2018) proposed a multimodal framework named EmotionMeter, which uses EEG and EYE signals to recognize four emotions. Liu et. al. (2021) used two multimodal emotion recognition models, deep canonical correlation analysis (DCCA) and bimodal deep autoencoder (BDAE), and conducted experiments on several multimodal emotion datasets. Jiang et. al. (2023) proposed a transformer-based model (MAET) to fuse the EEG and EYE signals, achieving good performance in the seven emotions classification task. All of these works have achieved good results, however, they require data from different modalities to be paired when testing or the models' performance will be greatly reduced or even unable to work.

Cross-modal Learning

Radford et. al. (2021) proposed the Contrastive Langauge-Image Pre-training (CLIP) model to achieve mutual conversion between text and image modalities, which is a classical work in the cross-modal field. More cross-modal work has been explored in the field of cross-modal video moment retrieval (Anne Hendricks et al. 2017; Gao et al. 2017; Shimomoto et al. 2022). Xia et. al. (2024) proposed a framework named Uni-Code to train a codebook and extract shared semantic information from paired multimodal data, such as video-audio-text.

As for the emotional BCI, Jiang et. al. (2019) proposed a BDAE-regressor method to predict multimodal features with EYE signals. Yan et. al. (2021) utilized a generative adversarial network (GAN) to generate multimodal features with EYE signals. Jiang et. al. (2024) proposed an EEG-assisted Contrastive Learning Framework with a Functional Emotion Transformer (ECO-FET) method for cross-modal emotion recognition. These methods simplify the modalities required for emotion recognition, but they only focus on EYE signals and do not take into account the generalization between any modalities.

Contrastive Learning

Contrastive learning (CL) is popular in self-supervised learning, as it allows models to learn intrinsic features of data that are independent of labels. Oord et. al. (2018) proposed Contrastive Predictive Coding (CPC), which is a highly generalized general framework. The core idea of CPC is to perform CL by predicting future embedding vectors. Its positive samples are the future embedding vectors obtained by the encoder from the future input, like x_{t+1}, and its negative samples are the embeddings corresponding to the input at any time. Tian et. al. (2020) proposed Contrastive Multiview Coding (CMC), whose core idea is that many perspectives of an object can be considered as positive samples. Yuan et. al. (2021) proposed a CL framework for multimodal alignment. All of these methods provide valuable ideas for our work.

Methodology

Multi-to-Single Emotion Recognition Model

The overall architecture and the pre-training phase of our model are shown in Figure 1(a). We design two encoders for each modality. One is a transformer-based emotion-related (ER) encoder, which is used to extract features about emotions from data, and the other is an MLP-based emotion-independent (EI) encoder to extract features unrelated to emotions. The ER encoder contains several transformer blocks consisting of $FeedForward$ and $LayerNormalization$ operations. For EEG and some other signals, we design a special STS attention, which is introduced in detail in the next subsection. To minimize the correlation between the features extracted by the two encoders, we use the Contrastive Log-ratio Upper Bound (CLUB) method (Cheng et al. 2020). Different from traditional methods such as InfoNCE, CLUB can minimize the upper bound of the mutual information, making two sets of data as independent as possible. Given any modality X, according to the variational CLUB term with conditional distribution un-

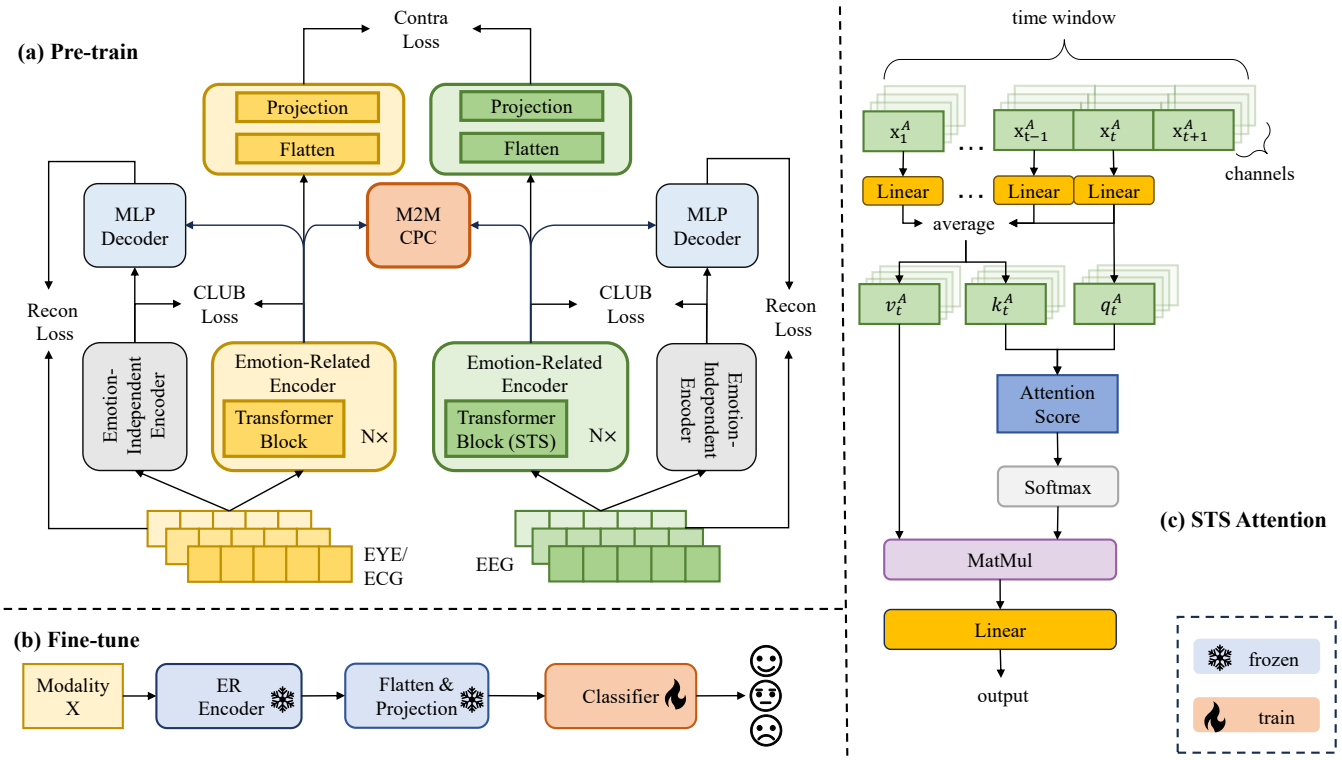

Figure 1: The overview of Multi-to-Single framework. (a) The pre-training phase of our model. The yellow and green modules represent the data and key structures corresponding to different modalities, respectively. (b) The fine-tuning phase of the model. In this stage, we input data from any modality X, freeze its corresponding encoder, and only optimize the newly added classifier. (c) The architecture of our proposed Spatial and Temporal-Sparse attention mechanism.

known, the CLUB loss of X is defined as:

$$\mathcal{L}^X_{CLUB}(z^X, z'^X) := \mathbb{E}_{p(z^X, z'^X)}[log(q_\theta(z'^X|z^X))] - \mathbb{E}_{p(z^X)}\mathbb{E}_{p(z'^X)}[log(q_\theta(z'^X|z^X))], \quad (1)$$

where z^X and z'^X are outputs of ER and EI encoders, respectively. $q_\theta(z'^X|z^X)$ is a variational distribution with parameter θ to approximate $p(z'^X|z^X)$.

To verify whether the encoders extract effective features of the data, we add a decoder for each modality and compute the reconstruction loss of each modality. The reconstruction loss of modality X is defined as:

$$\mathcal{L}^X_{Recon} = ||D_X(z^X, z'^X) - X||^2, \quad (2)$$

where D_X represents the decoder of modal X.

To learn and fuse features across different modalities, we use the contrastive learning method and propose a novel M2M CPC module. The M2M CPC is introduced in detail in the third subsection. For any embedding vectors z^X, it is projected into a new embedding space through an average pooling layer in the time window dimension and a linear layer. We define the final embedding of modality X as $\bar{z}^X \in \mathbb{R}^{S \times D_f}$, where S represents the number of samples and D_f denotes the final embedding dimension. There are S positive sample pairs and $S^2 - S$ negative sample pairs. According to the idea of InfoNCE, we can consider it as a S-class classification task. Taking EEG and EYE data as an example, we can define the ground truth GT as $[0, 1, \ldots, S-1]$ and compute the contrastive loss $\mathcal{L}_{Contra}$ as follows:

$$\mathcal{L}_{Contra} = (CrossEntropy(\bar{z}^{eeg}, GT) + CrossEntropy(\bar{z}^{eye}, GT))/2. \quad (3)$$

Spatial and Temporal-Sparse Attention Mechanism

In the experiment, we use many kinds of signals, including EEG, ECG, and PPC. These signals have the following commonality: they are all collected continuously for a long time through electrodes at certain fixed positions on the human body. Therefore, these signals have strong internal characteristics in both spatial and temporal dimensions. We propose a Spatial and Temporal-Sparse (STS) Attention Mechanism in our work to fully utilize this characteristic. The structure of STS Attention is shown in Figure 1(c). For data in any time window, in addition to learning the structural information between electrodes, we fuse the data in the temporal dimension. After linear transformation, we take the average of the values at time t, $t-1$, and the initial time as v^A_t and k^A_t.

This operation is based on the following considerations: firstly, the initial data within each time window determines the basic information of the data segment; secondly, the data

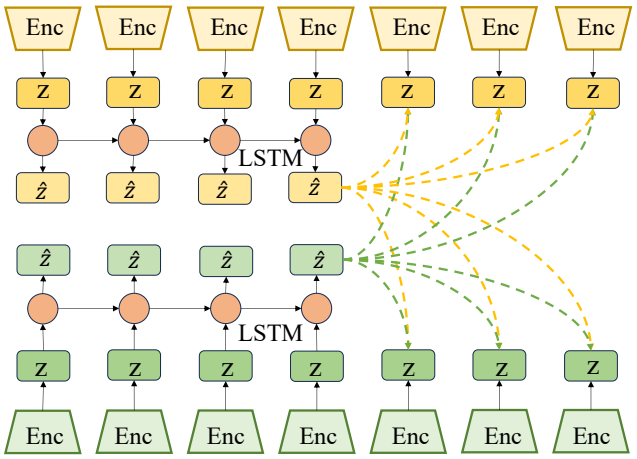

Figure 2: The structure of M2M CPC module. The yellow and green blocks represent different modalities.

at each moment is closely related to the previous moment. So now we have:

$$Q = W^Q x_t^A, \tag{4}$$

$$K = W^K(x_t^A \oplus x_{t-1}^A \oplus x_1^A), \tag{5}$$

$$V = W^V(x_t^A \oplus x_{t-1}^A \oplus x_1^A), \tag{6}$$

where $\oplus$ represents summation and average. Given attention dimension D_{att}, we can calculate the attention as follows:

$$Attention = softmax(\frac{QK^T}{\sqrt{D_{att}}})V. \tag{7}$$

Multi-to-Multi CPC

To learn the features between different modalities, we propose a novel Multi-to-Multi (M2M) CPC module, which is shown in Figure 2. The traditional CPC module extracts data features by predicting the correlation between future embedding vectors and current embedding vectors in a sequence using an autoregressive model. For human emotional changes, it often leads to simultaneous changes in multiple physiological signals of the human body. In practical situations, we can comprehensively judge the current emotional state of the subjects by detecting various physiological signal changes, which is an important reason why traditional multimodal models usually have better performance than unimodal models. So in the M2M CPC module, we use current embedding vectors from multiple modalities to predict future vectors of each modality.

Taking EEG and EYE data as an example, after passing through their emotion-related encoders, we get the embedding vectors:

$$z^{eeg} = (z_1^{eeg}, z_2^{eeg}, \ldots, z_T^{eeg}) \in \mathbb{R}^{T \times S \times D}, \tag{8}$$

$$z^{eye} = (z_1^{eye}, z_2^{eye}, \ldots, z_T^{eye}) \in \mathbb{R}^{T \times S \times D}, \tag{9}$$

where T represents the length of time window, S represents the number of samples of the data, and D denotes the embedding dimension. We use a two-layer LSTM as the autoregressive model for each modality. Define M as the length of observed sequence and N as the prediction step, we have:

$$\hat{z}^{eeg} = LSTM(z_1^{eeg}, \ldots, z_M^{eeg}) \in \mathbb{R}^{M \times S \times D'}, \tag{10}$$

$$\hat{z}^{eye} = LSTM(z_1^{eye}, \ldots, z_M^{eye}) \in \mathbb{R}^{M \times S \times D'}, \tag{11}$$

where D' is the hidden dimension of LSTM and $M + N \leq T$. Then we concatenate $\hat{z}_M^{eeg}$ and $\hat{z}_M^{eye}$ to predict the future N steps for each modality. We use InfoNCE loss to optimize the module. The loss function consists of four parts: the loss of predictive vectors and real vectors in the same modality and the loss of predictive vectors and real vectors in the other modality. Define Z_A and Z_B as sets of negative samples and one positive sample, where $A, B \in \{eeg, eye\}$. Then we have:

$$\mathcal{L}_{CPC}^{A2B} = -\frac{1}{N}\sum_{n=1}^{N} log\left[\frac{exp(z_{M+n}^B W_n^A[\hat{z}_M^A \cdot \hat{z}_M^B])}{\sum_{z_j \in Z_B} exp(z_j W_n^A[\hat{z}_M^A \cdot \hat{z}_M^B])}\right], \tag{12}$$

where $[\cdot]$ denotes concatenation operation. So the complete M2M CPC loss is:

$$\mathcal{L}_{CPC} = \mathcal{L}_{CPC}^{eeg2eye} + \mathcal{L}_{CPC}^{eeg2eeg} + \mathcal{L}_{CPC}^{eye2eeg} + \mathcal{L}_{CPC}^{eye2eye}. \tag{13}$$

Pre-train and Fine-tune

Taking into account all the losses mentioned above, we have obtained the final loss for the pre-training phase:

$$\mathcal{L} = \alpha'\mathcal{L}_{CLUB} + \beta'\mathcal{L}_{Recon} + \gamma'\mathcal{L}_{Contra} + \lambda'\mathcal{L}_{CPC}, \tag{14}$$

where α', β', γ', and λ' are corresponding coefficients. The complete pre-training process is shown in Algorithm 1.

In the fine-tuning stage, we only need to input data of one modality, freeze its encoder and linear projection layer, and add a classifier for optimization. For cross-modal tasks, we use different modalities to test. For unimodal tasks, the fine-tuning and testing process is similar to traditional supervised learning.

Algorithm 1: Pre-training Phase of M2S Model

Input: Unlabeled and paired data X^A, X^B from modalities A and B
Hyper Parameter: Learning rate and weight decay
Output: Pre-trained encoders

1: Input X^A, X^B into their corresponding encoders: $z^A = ER_A(X^A)$, $z'^A = EI_A(X^A)$; $z^B = ER_B(X^B)$, $z'^B = EI_B(X^B)$.
2: Compute $\mathcal{L}_{CLUB}^A(z^A, z'^A)$ and $\mathcal{L}_{CLUB}^B(z^B, z'^B)$.
3: Compute $\mathcal{L}_{Recon}^A$ and $\mathcal{L}_{Recon}^B$.
4: Input z^A and z^B into M2M CPC module.
5: Compute $\mathcal{L}_{CPC}^{A2B}$, $\mathcal{L}_{CPC}^{B2A}$, $\mathcal{L}_{CPC}^{A2A}$, and $\mathcal{L}_{CPC}^{B2B}$.
6: Flatten z^A and z^B, and project them into a new space.
7: Compute $\mathcal{L}_{Contra}$.
8: Optimize the final loss: $\mathcal{L} = \alpha'\mathcal{L}_{CLUB} + \beta'\mathcal{L}_{Recon} + \gamma'\mathcal{L}_{Contra} + \lambda'\mathcal{L}_{CPC}$.
9: **return** Encoders ER_A and ER_B.

Dataset	Model	EEG→EYE		EEG→EEG		EYE→EEG		EYE→EYE	
		Balanced Acc.	Cohen's Kappa	Balanced Acc.	Cohen's Kappa	Balanced Acc.	Cohen's Kappa	Balanced Acc.	Cohen's Kappa
SEED	BDAE-regressor	75.72/8.87	-	-	-	-	-	-	-
	BDAE-cGAN	81.02/8.04	-	-	-	-	-	-	-
	CLIP	85.23/11.99	77.76/18.04	89.51/11.70	84.24/17.56	80.65/10.84	70.99/16.22	88.42/8.75	82.53/13.22
	ECO-FET	72.56/8.42	58.69/12.71	88.44/11.32	82.64/16.96	73.79/8.53	60.54/12.91	82.50/12.18	73.58/18.42
	Uni-Code	69.44/14.72	53.83/22.31	74.01/12.65	60.79/19.19	74.11/10.46	61.02/15.81	69.76/15.50	54.30/23.50
	Multi-to-Single (Ours)	**91.72/9.53**	**87.50/14.41**	**93.18/8.19**	**89.75/12.30**	**91.92/9.34**	**86.85/14.90**	**93.65/7.99**	**90.44/12.02**
	w/o recon loss	86.90/10.34	80.34/15.48	91.17/9.95	86.67/15.06	86.38/10.54	79.46/15.90	86.95/9.90	80.34/14.93
	w/o cpc loss	83.90/10.36	75.77/15.60	90.31/9.81	85.40/14.80	90.54/8.48	85.77/12.79	89.04/9.26	83.49/13.93
	w/o club loss	87.63/9.52	81.35/14.37	90.22/11.65	85.26/17.58	88.13/11.33	82.14/17.09	86.99/10.52	80.37/15.90
	w/o contra loss	76.94/8.30	65.23/12.55	90.97/8.82	86.42/13.28	77.86/10.24	66.75/15.39	89.24/9.37	83.76/14.17
SEED-IV	BDAE-regressor	73.49/7.02	-	-	-	-	-	-	-
	BDAE-cGAN	75.74/6.66	-	-	-	-	-	-	-
	CLIP	77.73/11.38	64.75/19.49	79.13/12.30	65.95/22.70	75.10/11.70	57.37/25.08	81.54/11.14	70.87/19.65
	ECO-FET	70.03/11.29	54.81/16.79	80.46/14.71	69.20/24.27	68.14/12.77	50.27/20.31	81.23/13.77	71.05/22.02
	Uni-Code	66.23/14.71	47.34/24.98	61.53/14.99	38.51/25.70	60.86/13.75	36.14/24.97	65.92/13.82	48.88/21.27
	Multi-to-Single (Ours)	**84.70/11.35**	**75.58/19.59**	**85.47/11.05**	**76.95/19.10**	**81.90/11.49**	**71.29/19.59**	**87.31/10.52**	**79.85/18.67**
	w/o recon loss	78.41/12.99	67.53/19.15	83.66/11.19	73.85/19.36	78.89/12.40	67.04/20.23	84.32/11.41	76.04/18.49
	w/o cpc loss	78.61/13.33	66.93/21.42	80.99/12.19	69.38/22.36	78.98/10.04	65.02/23.60	81.25/13.29	70.93/21.46
	w/o club loss	82.34/11.72	72.66/19.15	79.31/11.53	67.49/19.39	79.09/12.80	66.42/22.35	83.00/11.22	73.33/18.87
	w/o contra loss	73.18/10.22	57.62/16.70	84.48/10.99	74.99/19.59	69.52/8.97	51.59/15.52	82.95/12.97	73.51/21.57
SEED-V	BDAE-regressor	72.80/5.07	-	-	-	-	-	-	-
	BDAE-cGAN	73.66/6.05	-	-	-	-	-	-	-
	CLIP	74.41/13.05	64.73/17.35	75.33/12.78	66.09/17.22	65.83/12.45	51.80/17.71	80.97/11.31	73.67/16.14
	ECO-FET	60.07/11.11	47.61/14.78	72.89/16.54	63.09/22.15	58.73/13.34	44.26/20.17	77.88/12.92	69.68/17.93
	Uni-Code	50.75/13.87	28.09/19.23	49.67/15.63	25.74/22.07	49.82/16.78	26.27/23.67	49.89/14.44	26.89/19.77
	Multi-to-Single (Ours)	**82.77/11.24**	**75.90/15.29**	**81.22/12.44**	**74.07/17.04**	**75.06/10.69**	**64.86/14.84**	**84.32//10.50**	**78.23/14.10**
	w/o recon loss	73.53/12.61	63.55/17.51	78.33/11.47	69.68/15.82	72.09/11.82	60.70/17.63	82.00/11.48	74.81/15.84
	w/o cpc loss	77.12/13.11	67.78/19.72	80.24/12.24	72.31/17.51	71.38/14.06	59.92/19.62	84.15/11.11	78.38/15.00
	w/o club loss	73.23/11.35	62.86/16.35	79.16/12.82	71.36/16.93	69.91/13.94	57.77/20.08	83.25/11.30	76.79/15.74
	w/o contra loss	71.27/12.50	60.61/16.27	79.71/11,71	71.93/16.06	66.31/11.88	53.73/15.41	82.49/11.35	75.56/15.78

Table 1: Subject-dependent accuracies (%) and kappa scores (%) for cross-modal and unimodal settings on SEED, SEED-IV, and SEED-V datasets. EEG → EYE represents fine-tuning with EEG and testing with EYE. The best results are in **bold** and the second-best results are underlined (excluding ablation results).

Experiments

Datasets and Experimental Details

In the experiment, we use five public multimodal emotion datasets: SEED (Duan, Zhu, and Lu 2013; Zheng and Lu 2015), SEED-IV (Zheng et al. 2018), SEED-V (Liu et al. 2021), DEAP (Koelstra et al. 2011), and DREAMER (Katsigiannis and Ramzan 2017). All five datasets use video stimuli to induce subjects' emotions. The basic information of these datasets is shown in Tabel 2. We extract DE features from EEG signals. Since the EEG signals in the downloaded DEAP dataset have already been filtered by a 4-75Hz filter, we only extract four frequency bands from the EEG signals in the DEAP, without the δ frequency band (1-3Hz). For the processing of ECG and PPS signals, we follow the processing method of Liu et. al. (2021) and extract 10-dimensional and 48-dimensional features, respectively. For EYE signals, we use 33-dimensional statistical features and computational features.

For the division of the training and testing sets, due to the fixed emotional labels corresponding to each video clip in the SEED series datasets, we divide the SEED, SEED-IV, and SEED-V datasets in ratios of 9:6, 16:8, and 10:5, respectively. The labels in the DEAP and DREAMER datasets are the scores of subjects on certain evaluation metrics, including valence, arousal, and dominance. This label leads to an uneven distribution of data, so we conduct four-fold and three-fold cross-validation on the DEAP and DREAMER, respectively. Each fold's training and testing ratios are 3:1 and 2:1, respectively.

Dataset (class)	Subjects	Videos	Modality/Dimension
SEED (3)	12 (x3)	15	EEG/310 (62 * 5) EYE/33
SEED-IV (4)	15 (x3)	24	EEG/310 (62 * 5) EYE/31
SEED-V(5)	16 (x3)	15	EEG/310 (62 * 5) EYE/33
DEAP (2)	32	40	EEG/128 (32 * 4) PPS/48 (8 * 6)
DREAMER (2)	23	18	EEG/70 (14 * 5) ECG/10 (2 * 5)

Table 2: The basic information of five public multimodal emotion datasets. $Subjects$ represents the number of subjects, $(\times 3)$ indicates that each subject conducts three experiments. $Vidoes$ represents the number of video clips used in the dataset. The last column contains the modalities in each dataset and their data dimensions, the content in the brackets represents channels $\times$ frequency bands/statistical features.

We choose balanced accuracies and kappa scores as the classification evaluation metrics for the SEED series dataset. Due to individual subjects in DEAP and DREAMER datasets having only one category of label in a certain fold, it is hard to calculate kappa scores for every subject. Therefore, we calculate the balanced accuracies and F1 scores for these two datasets.

Dataset	Label	Model	EEG → PPS		EEG → EEG		PPS → EEG		PPS → PPS	
			Balanced Acc.	F1 Score	Balanced Acc.	F1 Score	Balanced Acc.	F1 Score	Balanced Acc.	F1 Score
DEAP	Valance	CLIP	65.79/16.65	59.51/19.84	65.83/16.07	60.29/18.63	64.71/16.07	58.99/18.67	66.91/16.42	60.89/19.11
		ECO-FET	71.62/12.08	54.94/15.31	70.15/13.83	62.15/15.34	73.96/12.37	53.99/14.57	67.74/15.86	61.34/17.27
		Uni-Code	68.19/11.32	57.81/14.86	66.54/12.07	55.54/15.53	67.01/12.66	57.24/15.81	68.45/11.38	58.25/13.33
		M2S (Ours)	**78.78/10.51**	**73.30/12.96**	**75.03/10.82**	**68.63/13.63**	**75.74/10.87**	**69.81/12.96**	**79.39/10.12**	**73.00/12.63**
	Arousal	CLIP	67.13/15.62	61.55/21.68	66.40/16.15	60.54/18.08	65.75/16.01	59.88/17.40	68.08/15.64	61.66/17.78
		ECO-FET	75.26/8.72	53.87/14.39	70.08/15.63	61.47/16.70	77.76/9.50	52.97/12.87	67.38/15.19	60.28/16.56
		Uni-Code	69.39/9.90	53.69/12.75	70.00/12.74	54.92/14.28	69.03/12.94	54.22/15.86	69.30/10.12	52.49/12.96
		M2S (Ours)	**81.62/10.64**	**72.51/14.85**	**77.24/12.07**	**68.34/15.51**	**78.62/14.85**	**69.90/14.73**	**80.93/10.76**	**70.71/15.02**
Dataset	Label	Model	EEG → ECG		EEG → EEG		ECG → EEG		ECG → ECG	
			Balanced Acc.	F1 Score	Balanced Acc.	F1 Score	Balanced Acc.	F1 Score	Balanced Acc.	F1 Score
DREAMER	Valance	CLIP	85.06/11.05	82.17/11.71	86.38/11.24	82.78/13.86	89.36/10.60	86.13/13.05	86.90/10.61	83.30/13.57
		ECO-FET	67.07/8.06	58.88/9.04	87.39/10.26	83.64/13.52	67.74/7.51	59.82/9.86	83.90/12.56	79.36/16.93
		Uni-Code	64.29/15.36	53.44/21.68	63.04/15.38	51.40/22.17	61.87/15.86	49.81/24.30	63.04/15.67	52.52/22.19
		M2S (Ours)	**90.27/10.82**	**88.31/17.66**	**91.71/8.75**	**90.28/13.30**	**91.56/8.59**	**90.12/13.25**	**89.32/10.30**	**87.76/16.23**
	Arousal	CLIP	89.28/10.51	87.00/12.85	90.46/11.53	86.73/15.24	91.82/9.20	88.54/12.01	90.81/9.70	87.72/13.19
		ECO-FET	70.08/10.05	57.14/11.05	92.07/8.89	86.41/15.08	73.11/11.12	66.79/11.68	89.99/8.56	83.16/15.28
		Uni-Code	67.70/20.08	56.81/25.36	69.68/21.21	60.17/26.59	70.21/20.60	62.18/25.62	67.69/19.96	57.17/26.25
		M2S (Ours)	**93.52/7.23**	**94.23/7.29**	**95.16/5.39**	**94.93/6.20**	**95.66/5.91**	**95.91/5.54**	**94.54/6.59**	**94.52/6.87**
	Dominance	CLIP	88.42/10.16	83.21/14.44	90.35/9.80	86.03/12.91	91.37/9.72	87.89/13.11	87.16/12.29	82.41/15.70
		ECO-FET	70.20/7.91	57.13/11.34	88.44/10.06	83.10/14.97	70.77/7.31	58.48/8.89	87.94/11.52	83.03/15.24
		Uni-Code	68.17/19.27	57.59/25.29	66.81/18.86	58.51/24.77	68.04/19.14	59.88/24.68	67.43/19.22	63.20/24.84
		M2S (Ours)	**92.74/6.80**	**93.05/7.03**	**94.29/6.82**	**94.63/6.75**	**94.26/6.4**	**94.39/6.58**	**91.39/7.06**	**91.45/7.87**

Table 3: Subject-dependent accuracies (%) and F1 scores (%) for cross-modal and unimodal settings on DEAP and DREAMER datasets. The best results are in **bold** and the second-best results are underlined.

Baselines

We choose several state-of-the-art cross-modal methods as baselines, including BDAE-regressor (Jiang et al. 2019), BDAE-cGAN (Yan, Zhao, and Lu 2021), ECO-FET (Jiang et al. 2024), CLIP, and Uni-Code (Xia et al. 2024). We use the same encoders and fine-tuning process as our model for CLIP to ensure fairness in comparison. We also compare some classical and state-of-the-art supervised multimodal methods, including BDAE (Liu, Zheng, and Lu 2016), Emotion Transformer Fusion (ETF) (Wang et al. 2021), VigilanceNet (Cheng et al. 2022), and MAET (Jiang et al. 2023) models, to determine whether the pre-training of the model and the contrastive learning modules play a role in learning and fusing key information between different modalities.

Results

Compared to Cross-modal Methods Cross-modal and unimodal experimental results of SEED series datasets are shown in Tabel 1 and the results of DEAP and DREAMER datasets are shown in Tabel 3. As shown in the tables, our model outperforms all the previous work in various settings. Both cross-modal and unimodal tasks have significantly improved accuracies. In all cross-modal tasks, our model generally improves by over 5 percentage points compared to existing suboptimal methods, with a maximum improvement of over 10 percentage points (Independent Samples t-test, $p < 0.05$). After using a certain modality for fine-tuning, the test performance of single-modality is slightly better than that of cross-modality.

Compared to Multimodal Methods The results compared to supervised multimodal methods are shown in Table 4. Since EEG is still the most concerning data in most studies, we only present the results of fine-tuning on EEG modality for our model. As shown in the table, with only one two-layer MLP as the classifier for fine-tuning, our model outperforms most of the supervised multimodal methods. At the same time, we test the unimodal supervised training of the ER encoder that only uses EEG. Its accuracy is slightly lower than using the pre-trained model, but can still achieve the performance of some multimodal methods. The experimental results indicate that our pre-training process plays a role and the introduction of contrastive learning modules has also enabled the effective fusion of features from differ-

Dataset	Model	Balanced Acc.	Cohen's Kappa
SEED	BDAE	87.00/10.83	80.44/16.27
	ETF	89.45/9.68	84.10/14.60
	VigilanceNet	87.67/11.67	81.41/17.63
	MAET	**93.38/9.29**	**90.01/14.05**
	M2S (S)	87.50/9.80	81.14/14.82
	M2S	93.18/8.19	89.75/12.30
SEED-IV	BDAE	83.13/12.63	73.86/21.10
	ETF	80.31/12.81	70.02/20.62
	VigilanceNet	78.23/13.63	67.46/20.66
	MAET	83.81/15.00	76.17/22.48
	M2S (S)	83.26/12.43	74.32/18.54
	M2S	**85.47/11.05**	**76.95/19.10**
SEED-V	BDAE	71.79/15.93	62.71/20.77
	ETF	77.29/12.84	68.66/17.69
	VigilanceNet	69.39/13.79	58.25/19.20
	MAET	76.53/14.66	68.19/19.83
	M2S (S)	72.14/13.84	60.82/20.79
	M2S	**81.22/12.44**	**74.07/17.04**

Table 4: Balanced accuracies (%) and kappa scores (%) of multimodal methods compared with M2S on SEED, SEED-IV, and SEED-V datasets. (S) indicates that the same encoder is used for unimodal supervised learning.

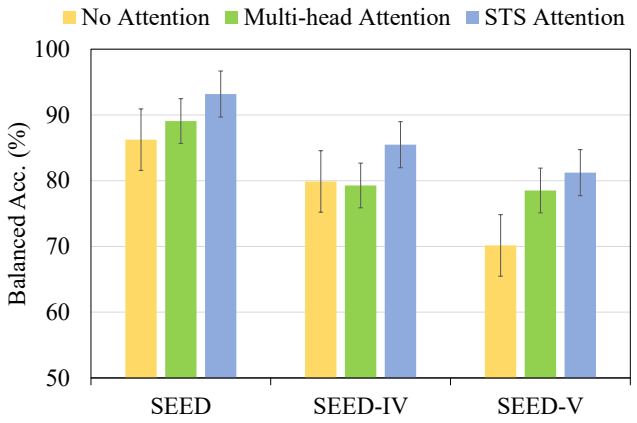

Figure 3: The results of different attention mechanisms on SEED, SEED-IV, and SEED-V datasets based on EEG unimodal fine-tuning.

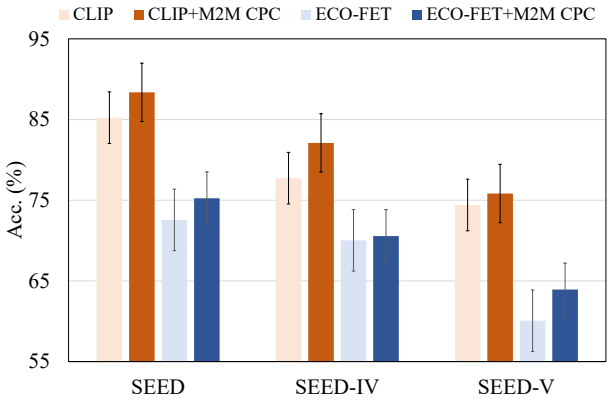

Figure 4: Cross-modal task (EEG→EYE) accuracy changes of CLIP and ECO-FET models on SEED, SEED-IV, and SEED-V datasets after adding M2M CPC module.

ent modalities.

Ablation Studies

The Effect of Different Loss Functions As shown in Table 1, we conduct ablation experiments on each of the four loss functions in the model. After removing a certain loss function, the performance of the model decreases to varying degrees. Removing the loss function of contrastive learning has the most severe impact on cross-modal tasks.

The Effect of STS Attention Mechanism To verify the effectiveness of the STS attention mechanism, we conduct the following experiments on SEED series datasets: replacing the STS attention mechanism with a regular multi-head attention mechanism or directly removing the attention mechanism. Since the STS attention mechanism is only applied to signals collected by electrodes, we only present the results of EEG unimodal fine-tuning. As shown in Figure 3, the performance of the model has been significantly improved with the STS attention mechanism.

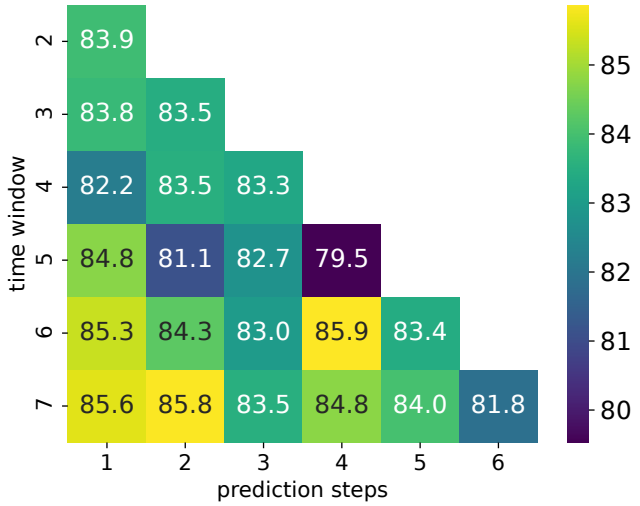

Figure 5: The heatmap of the cross-modal task (EEG→EYE) on the SEED dataset. The value of the unit is balanced accuracy (%). The higher the accuracy, the brighter the color of the unit.

Ablation Study of M2M CPC Module To analyze the impact of the M2M CPC module, we first test the performance of existing methods after adding this module. The results are shown in Figure 4. After adding M2M CPC, each modality learns each other's features through predictive coding, achieving feature fusion and improving accuracies. We also test the effect of different lengths of time windows and prediction steps on the final accuracy. We plot heat maps for two cross-modal tasks on the SEED dataset. As shown in Figure 5, within the same hyperparameter tuning range, when the time window length is fixed, the model's accuracy generally shows an upward trend followed by a downward trend as the number of prediction steps increases. When the prediction step length is fixed, a longer time window helps the model perform better.

Conclusion

In this paper, we propose a novel method M2S for cross-modal learning, which can reduce multimodal dependency in emotion recognition and achieve multimodal performance using only a single modality. We conduct various experiments on multiple datasets, and the experimental results show that our model achieves the SOTA performance in the cross-modal tasks. Furthermore, our proposed M2M CPC module can be generalized to other models to learn and fuse features from different modalities and help existing models perform better.

Acknowledgements

This work was supported in part by grants from National Natural Science Foundation of China (Grant No. 62376158), STI 2030-Major Projects+2022ZD0208500, Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZD ZX), Shanghai Pujiang Program (Grant

No. 22PJ1408600), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25, YG2024ZD25 and YG2024QNA03), Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University (No. 21TQ1400203), GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine, and Shanghai Jiao Tong University SEIEE-Shanghai Emotionhelper Technology Co., Ltd Joint Laboratory of Affective Brain-Computer Interfaces.

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, 5803–5812.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, 1779–1788. PMLR.

Cheng, X.; Wei, W.; Du, C.; Qiu, S.; Tian, S.; Ma, X.; and He, H. 2022. Vigilancenet: decouple intra-and inter-modality learning for multimodal vigilance estimation in RSVP-based BCI. In *Proceedings of the 30th ACM International Conference on Multimedia*, 209–217.

Duan, R.-N.; Zhu, J.-Y.; and Lu, B.-L. 2013. Differential entropy feature for EEG-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 81–84. IEEE.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, 5267–5275.

Jiang, H.; Guan, X.; Zhao, W.-Y.; Zhao, L.-M.; and Lu, B.-L. 2019. Generating Multimodal Features for Emotion Classification from Eye Movement Signals. *Aust. J. Intell. Inf. Process. Syst.*, 15(3): 59–66.

Jiang, W.-B.; Li, Z.; Zheng, W.-L.; and Lu, B.-L. 2024. Functional emotion transformer for EEG-assisted cross-modal emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1841–1845. IEEE.

Jiang, W.-B.; Liu, X.-H.; Zheng, W.-L.; and Lu, B.-L. 2023. Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5975–5984.

Katsigiannis, S.; and Ramzan, N. 2017. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1): 98–107.

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

Lan, X.; Yuan, Y.; Wang, X.; Wang, Z.; and Zhu, W. 2023. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2): 1–33.

Li, X.; Zhang, Y.; Tiwari, P.; Song, D.; Hu, B.; Yang, M.; Zhao, Z.; Kumar, N.; and Marttinen, P. 2022. EEG based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4): 1–57.

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

Liu, W.; Zheng, W.-L.; and Lu, B.-L. 2016. Emotion recognition using multimodal deep learning. In *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part II 23*, 521–529. Springer.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *ArXiv preprint arXiv:1807.03748*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Shimomoto, E. K.; Marrese-Taylor, E.; Takamura, H.; Kobayashi, I.; Nakayama, H.; and Miyao, Y. 2022. Towards parameter-efficient integration of pre-trained language models in temporal video grounding. *ArXiv preprint arXiv:2209.13359*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.

Wang, Y.; Jiang, W.-B.; Li, R.; and Lu, B.-L. 2021. Emotion transformer fusion: Complementary representation properties of EEG and eye movements on recognizing anger and surprise. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1575–1578. IEEE.

Xia, Y.; Huang, H.; Zhu, J.; and Zhao, Z. 2024. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36.

Yan, X.; Zhao, L.-M.; and Lu, B.-L. 2021. Simplifying multimodal emotion recognition with single eye movement modality. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1057–1063.

Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; and Faieta, B. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6995–7004.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2023. Temporal sentence grounding in videos: A survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10443–10465.

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

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