Adversarial Training based Domain Adaptation for Cross-Subject Emotion Recognition

Sungpil Woo*
ETRI / KAIST
Daejeon, South Korea
woosungpil@etri.re.kr

Muhammad Zubair*
ETRI
Daejeon, South Korea
zubair5608@etri.re.kr

Sunhwan Lim
ETRI
Daejeon, South Korea
shlim@etri.re.kr

Daeyoung Kim
KAIST
Daejeon, South Korea
kimd@kaist.ac.kr

Abstract

Emotion recognition using electroencephalography (EEG) signals faces challenges due to inter-subject variability. While domain adaptation techniques have been explored, achieving domain invariant and emotion-specific feature extraction remains difficult. To address this, we propose the adversarial training based domain adaptive representation learning method. The proposed model focuses on learning domain invariant and emotion-specific features for a generalized emotion recognition model across subjects. It uses a deep architecture with channel-wise and feature-wise attention mechanisms to separate emotion-specific and domain-specific features. These features assist the domain discriminator in learning generalized EEG representations. Experiments on SEED show that our model significantly improves emotion recognition and mitigates domain shift by learning generalized EEG signal representations.

1 Introduction

Emotions are complex psycho-physiological responses that are essential to human cognition[16, 35, 37]. In the field of Human-Computer Interaction (HCI), emotion recognition plays a crucial role by enabling computers to understand human emotional states[5] and assisting healthcare professionals in comprehending the behavior of mental patients[2]. Emotion recognition methods are generally categorized into two types: those based on physical signals (such as speech, body posture, and facial expressions) [33, 23]and those based on physiological signals (such as EEG, ECG, and GSR)[12, 9]. Among these, EEG signals, which reflect brain activity with high temporal resolution, are regarded as the most reliable for recognizing true emotions.

Despite extensive research, EEG-based emotion recognition faces significant challenges. One major issue is the selective reflection of brain activity by different channels and the presence of irrelevant features, which can degrade classification performance[38, 14]. Therefore, the automatic selection of relevant EEG channels and features is crucial. Another significant challenge is the data distribution shift caused by the non-stationary nature of EEG signals and inter-subject variability[6, 13], which limits model generalizability.

To address these challenges, we propose the Adversarial Training based Domain Adaptation method. Our model learns domain-invariant and emotion-specific features by using an attention mechanism to

*equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

highlight relevant EEG characteristics. It employs adversarial domain adaptation to align features across different subjects. Experiments conducted on the SEED dataset demonstrate that our model significantly enhances emotion recognition and mitigates domain shift.

2 Related Work

Recent research on emotion recognition using EEG signals has shifted from traditional machine learning methods, like SVM[1], k-NN[11], LDA[27], QDA[14], and Naive Bayes[34], to deep learning techniques. Deep learning models, including 1D, 2D, and 3D CNNs[8], LSTMs[6, 24, 32], and GCNNs[26], have shown significant improvements by capturing spatio-temporal dependencies and relationships between EEG channels. Additionally, VAEs[30] and GANs[36] have been employed to enhance emotion classification performance.

Attention mechanisms have been incorporated into deep learning models to address the challenge of selecting relevant EEG channels and features[3, 21, 31]. These mechanisms adaptively highlight vital information, improving the extraction of discriminative features[29]. CNNs, LSTMs, and GCNNs have all been enhanced with attention modules to refine spatial and temporal feature extraction, boosting classifier performance.

EEG signals vary between individuals, causing distribution shifts that hinder model generalization. Recent studies have applied domain adaptation techniques to mitigate this issue, using methods like domain discriminators, joint distribution adaptation, and adversarial inference[6, 19, 20]. Solutions addressing both global and local distribution shifts[17, 18], such as multi-stage adaptation and center loss, have also been proposed to improve model robustness across different subjects.

3 Methodology

The proposed method aims to extract emotion-specific and domain-invariant features from EEG signals to enhance model generalization. The approach involves an adversiral learning based domain adaptation strategy, which aligns features to mitigate distribution shift problems, improving classification performance.

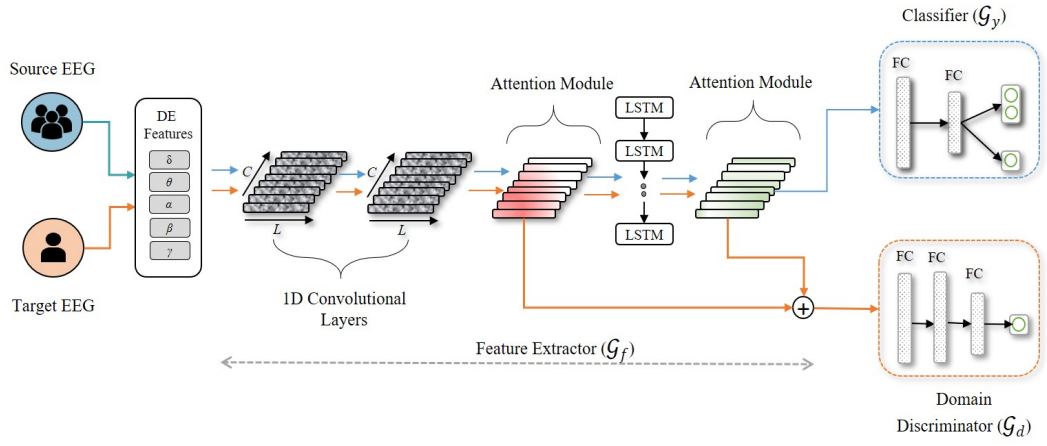

Figure 1: The proposed method for EEG-based emotion recognition includes (1) a feature extractor $\mathcal{G}_f$ integrated with feature and channel attention module; (2) a classifier $\mathcal{G}_y$; (3) a domain discriminator $\mathcal{G}_d$. The proposed deep model takes the differential entropy (DE) features of all bands in each EEG channel as input. The attention modules highlight the emotion-specific features and provide a domain discriminator with a set of peripheral features $\hat{f}_{c,l}$. The domain discriminator adversarially optimizes the feature extractor to project the target and source data features into a unified feature space.

3.1 Model Architecture

The proposed model consists of a feature extractor (CNN and LSTM layers), a classifier, and a domain discriminator, integrated with attention modules. The feature extractor learns inter-relationships between EEG channels, while the attention modules highlight emotion-specific and domain-invariant features.

Feature Extractor uses CNN and LSTM layers to learn inter-relationships between EEG channels and extract high-level representations. Attention modules refine these features, highlighting the most relevant emotion-specific channels and features.

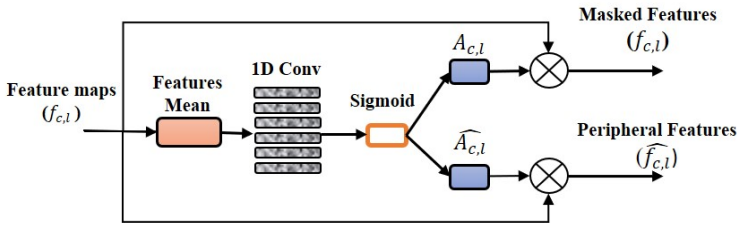

Figure 2: Proposed attention module with the most relevant and peripheral set of features.

The attention module scales each feature or channel based on its significance to the target objective, emphasizing task-relevant information. It generates both target-specific and peripheral features, aiding the domain discriminator in handling data distribution shifts.

The feature classifier uses fully connected layers to map features to emotion classes, utilizing cross-entropy loss to enhance discriminative feature learning.

The domain discriminator, composed of dense layers, aligns source and target domain features using adversarial training, encouraging the feature extractor to learn domain-invariant representations. Peripheral features from attention modules further assist in this alignment.

3.2 Adversarial Training based Domain Adaptation Strategy

The strategy involves pre-training the model on source data and employing adversarial training to align source and target feature distributions. The loss functions used include cross-entropy and binary cross-entropy for domain discrimination.

Algorithm 1 Adversarial training based domain adaptation

Input: Source dataset $S = \{(x_i^s, y_i^s)\}$; Target dataset $T = \{(x_i^t)\}$ A pre-trained model $\mathcal{M}(x_s, \mathcal{G}_f, \mathcal{G}_y)$. Learning rate λ and number of iterations N.
Output: Domain-invariant model $\mathcal{M}(\theta_f, \theta_c, \theta_d)$

1: **for** $i = 0$ **to** N **do**
2: Using (x^s, y^s), compute emotion-specific and peripheral features $\{f_{c,l}^m, \hat{f}_{c,l}^m\}$
3: Compute $\mathcal{L}_{ce}$
4: Update parameters of feature extractor $\mathcal{G}_f(\theta_f)$ and classifier $\mathcal{G}_y(\theta_c)$
$\theta_f \leftarrow \theta_f - \lambda \frac{\partial \mathcal{L}_{ce}}{\partial \theta_f}$, $\theta_c \leftarrow \theta_c - \lambda \frac{\partial \mathcal{L}_{ce}}{\partial \theta_c}$
5: Compute $\mathcal{L}_{adv}$ using $\{(x^s, \hat{y}^s), (x^t, \hat{y}^t)\}$ and $\{\hat{f}_{c,l}^m\}$
6: Update parameters of feature extractor $\mathcal{G}_f(\theta_f)$
$\theta_f \leftarrow \theta_f - \lambda \frac{\partial \mathcal{L}_{adv}}{\partial \theta_f}$
7: Compute $\mathcal{L}_{dis}$ using $\{(x^s, y^s), (x^t, y^t)\}$ and $\{\hat{f}_{c,l}^m\}$
8: Update parameters of domain discriminator $\mathcal{G}_d(\theta_d)$
$\theta_d \leftarrow \theta_d - \lambda \frac{\partial \mathcal{L}_{dis}}{\partial \theta_d}$
9: **end for**
10: Return $\mathcal{M}(\theta_f, \theta_c, \theta_d)$

The $\mathcal{L}_{ce}$ is cross-entropy loss and optimized over $\mathcal{G}_f^s$ and $\mathcal{G}_y$ using labeled data from source dataset. This optimization assists the model in classifying emotion classes. For the domain prediction task, the domain discriminator is trained using binary cross-entropy loss function $\mathcal{L}_{dis}$. For domain discrimination, source labels are assigned as 0, and target labels are assigned as 1. The loss function $\mathcal{L}_{dis}$ optimizes the domain discriminator module $\mathcal{G}_d$ to determine whether the sample belongs to the source data or the target data.

Furthermore, adversarial training is performed to match the feature distribution of the source and target domain. In this regard, the parameters of the feature extractor $\mathcal{G}_f$ are optimized to learn a domain-invariant representation that projects source and target domain features into a unified features space for efficient classification. During adversarial training, the domain labels of source and target data are flipped to confuse the feature extractor to distinguish between the source and target domain. For adversarial loss $\mathcal{L}_{adv}$, we employed binary cross-entropy loss as given below.

$$\mathcal{L}_{adv} = \alpha \times \mathcal{L}_{ce}(\hat{y}_s, \hat{y}_t) \tag{1}$$

where α is a hyper-parameter that controls the impact of adversarial training, $\hat{y}_s$, $\hat{y}_t$ are the flipped labels of corresponding source and target data samples. The above loss function optimizes the feature extractor's parameters during adversarial training and enforces the model to learn a domain invariant representation for alleviating the problem of distribution shift in EEG-based emotion recognition.

The approach is detailed in the provided algorithm 1, outlining steps for computing losses, updating model parameters, and optimizing the domain-invariant model for efficient classification of data from different individuals.

4 Experiments

4.1 Dataset

Publically available datasets, namely SEED[21, 7] is used to evaluate the proposed adversarial training based domain adaptation method. SEED includes EEG recordings for three emotions (positive, neutral, negative) elicited by 15 movie clips, recorded from 15 subjects. Data is downsampled to 200Hz and a bandpass filter (0-75Hz) is applied to eliminate different noises and artifacts.

4.2 Feature Extraction

Differential entropy is used as the feature extraction method, breaking EEG signals into five frequency bands (delta, theta, alpha, beta, gamma) and computing differential entropy for each band in each channel. Following that, we estimate differential entropy for each of these bands over all channels as follows:

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\pi\sigma^2} \log \frac{1}{\sqrt{2\pi\sigma^2}} \tag{2}$$

$$exp \frac{(x-\mu)^2}{2\pi\sigma^2} = \frac{1}{2} \log 2\pi e\sigma^2 \tag{3}$$

We computed differential entropy using a 1-second long window. For each band in each channel, the differential entropy is computed to generate an input array with a size of (62, 5).

4.3 Experimental Settings

Experiments are conducted under subject-independent evaluation paradigms. In the subject-independent evaluation, leave-one-subject-out cross-validation (LOSOCV) is used, with the data of 14 subjects for training and the left-out subject for testing.

5 Results and Discussion

5.1 Evaluation on SEED

The proposed emotion recognition method is evaluated on the SEED dataset using subject-independent paradigms. The SEED dataset includes three distinct emotion classes: positive, neutral, and negative, which were used for classification in this study. Table 1 presents the experimental results of the proposed method on the SEED dataset. Compared to the baseline model, the proposed model achieves higher classification accuracy, demonstrating its generalization capability with an accuracy improvement from 85.76% to 91.28%.

Table 1: Classification Performance on SEED Dataset

Method	Average Accuracy (± Std.)
Base model	85.76% (±7.80)
Proposed model	**91.28% (±6.45)**

5.2 Visualizations

This section provides a visual illustration of the proposed model's effectiveness in recognizing human emotions. The objectives of these studies include the selection of emotion-specific information and the calibration of features across different domains. To achieve these goals, a channel and feature attention module, along with a domain discriminator, are incorporated. Additionally, an adversarial training method is proposed to assist the domain discriminator in learning a domain-invariant representation. The following subsections demonstrate the potential of these components.

5.2.1 Attention

To evaluate the attention modules' effectiveness, we visualized the channel attention masks for 62 electrode channels from the SEED dataset samples. These topographic maps show the correlation between channel activation and brain regions. Literature indicates negative emotions link to the brain's right hemisphere and positive emotions to the left[15, 1]. These maps confirm previous findings and validate that the proposed channel attention mechanism effectively highlights the most relevant channels for emotion classification.

Similarly, selecting relevant features and constraining irrelevant ones is crucial for EEG-based emotion classification[28, 14]. The proposed feature attention module focuses on the most contributing features. Heat maps of feature attention scores for SEED samples illustrate that the module assigns varying ratings to each feature based on its contribution to classification and relevance to the target emotion.

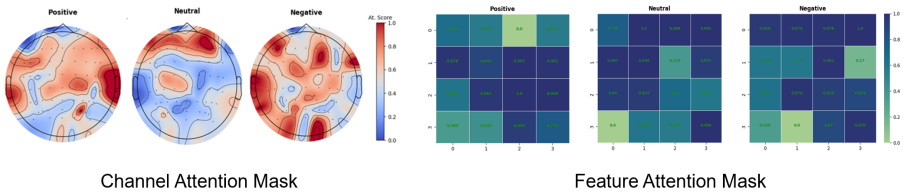

Figure 3: The Visualization of Channel attention(Left) and Feature attention(Right) masks

5.2.2 Representation Learning

This study uses domain adaptation to address inter-subject variation, enhancing classification accuracy and model generalization by aligning features across different domains. Comparisons between models trained with and without domain adaptation show that the proposed model significantly aligns source and target domain features, capturing intrinsic and domain-invariant features and alleviating intra-source distribution shifts as shown by t-SNE visualizations of SEED dataset.

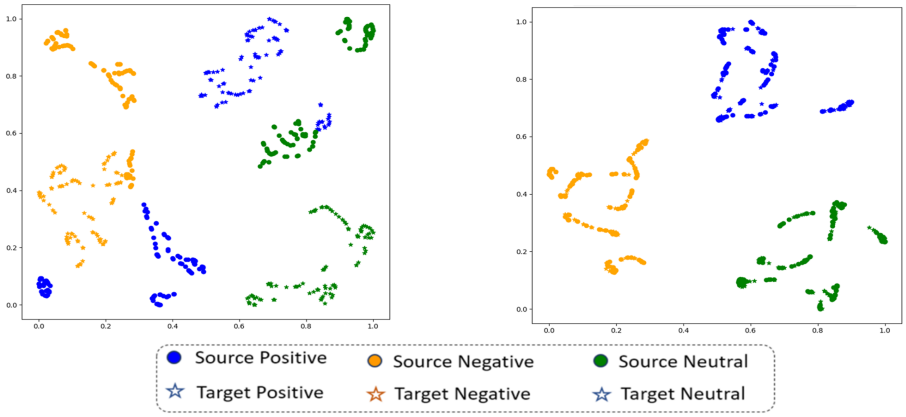

Figure 4: T-SNE visualization of the learned feature representations of Source and Target domain for the base model(Left) and the proposed model(Right)

5.3 Comparison with Previous Work

We compared the proposed method with recent studies on EEG-based emotion recognition using domain adaptation, focusing on those reporting classification performance on the SEED dataset. Table 2 summarizes these studies. Most studies focus only on feature selection and treat all channels uniformly. In contrast, our method uses an attention mechanism to highlight relevant features and channels. Unlike other studies [17, 19, 22, 10, 4, 25], the proposed domain adaptation method is assisted by the peripheral features (domain-specific features) extracted by attention mechanisms to accelerate feature calibration between source and target domains.

Table 2: Classification comparison with previously published results on SEED

Authors (Year)	Average Accuracy (± Std.)
Jinpeng Li et al.[17] (2019)	88.28% (±11.44)
Xiaobing Du et al.[6] (2020)	90.92% (±1.05)
Hao Chen1 et al.[4] (2021)	89.63% (±6.79)
Zhunan Li et al.[19] (2022)	91.08% (±7.70)
Qingshan She et al.[25] (2023)	86.16% (±7.87)
Proposed	**91.28% (±6.45)**

6 Conclusion

This paper proposes a adversarial traning based domain adaptation method to recognize emotion using EEG signals. To highlight the most contributing and relevant features and channels (EEG electrodes), an attention mechanism is introduced in this study. In addition to target-specific information, the proposed attention module also highlights peripheral features representing domain-specific information. Additionally, to reduce the distribution shift between source and target data, we adopted a adversarial learning method governed by a novel supervision strategy. The experimental results demonstrate that the proposed model can significantly learn a domain-invariant representation of EEG signals for classifying different emotions. The reported results also validate the effectiveness of the proposed attention mechanism in refining emotion-specific information for improved classification performance. Future work on EEG-based emotion recognition should concentrate on handling distribution shifts in source data adversarially. Moreover, further investigation on the connectivity between human emotion and EEG channels (electrodes) would be of great interest.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. RS-2024-00423362, 50%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-01032, Development of Collective Collaboration Intelligence Framework for Internet of Autonomous Things, 50%)

References

[1] S. M. Alarcao and M. J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2017.

[2] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya. Eeg-based emotion recognition approach for e-healthcare applications. In *2016 eighth international conference on ubiquitous and future networks (ICUFN)*, pages 946–950. IEEE, 2016.

[3] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad. A review of channel selection algorithms for eeg signal processing. *EURASIP Journal on Advances in Signal Processing*, 2015:1–21, 2015.

[4] H. Chen, M. Jin, Z. Li, C. Fan, J. Li, and H. He. Ms-mda: Multisource marginal distribution adaptation for cross-subject and cross-session eeg emotion recognition. *Frontiers in Neuroscience*, 15:778488, 2021.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[6] X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu, and H. Wang. An efficient lstm network for emotion recognition from multichannel eeg signals. *IEEE Transactions on Affective Computing*, 13(3):1528–1540, 2020.

[7] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pages 81–84. IEEE, 2013.

[8] Z. Gao, X. Wang, Y. Yang, Y. Li, K. Ma, and G. Chen. A channel-fused dense convolutional network for eeg-based emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 13(4):945–954, 2020.

[9] A. Goshvarpour, A. Abbasi, and A. Goshvarpour. An accurate emotion recognition system using ecg and gsr signals and matching pursuit method. *Biomedical journal*, 40(6):355–368, 2017.

[10] W. Guo, G. Xu, and Y. Wang. Multi-source domain adaptation with spatio-temporal feature extractor for eeg emotion recognition. *Biomedical Signal Processing and Control*, 84:104998, 2023.

[11] S. K. Hadjidimitriou and L. J. Hadjileontiadis. Toward an eeg-based recognition of music liking using time-frequency analysis. *IEEE Transactions on Biomedical Engineering*, 59(12):3498–3510, 2012.

[12] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung. Automatic ecg-based emotion recognition in music listening. *IEEE Transactions on Affective Computing*, 11(1):85–99, 2020.

[13] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

[14] R. Jenke, A. Peer, and M. Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective computing*, 5(3):327–339, 2014.

[15] B. H. Kim and S. Jo. Deep physiological affect network for the recognition of human emotions. *IEEE Transactions on Affective Computing*, 11(2):230–243, 2018.

[16] P. R. Kleinginna Jr and A. M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379, 1981.

[17] J. Li, S. Qiu, C. Du, Y. Wang, and H. He. Domain adaptation for eeg emotion recognition based on latent representation similarity. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):344–353, 2019.

[18] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou. A bi-hemisphere domain adversarial neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*, 12(2):494–504, 2018.

[19] Z. Li, E. Zhu, M. Jin, C. Fan, H. He, T. Cai, and J. Li. Dynamic domain adaptation for class-aware cross-subject and cross-session eeg emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 26(12):5964–5973, 2022.

[20] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş. Learning invariant representations from eeg via adversarial inference. *IEEE access*, 8:27074–27085, 2020.

[21] M. S. Özerdem and H. Polat. Emotion recognition based on eeg features in movie clips with channel selection. *Brain informatics*, 4(4):241–252, 2017.

[22] J. Quan, Y. Li, L. Wang, R. He, S. Yang, and L. Guo. Eeg-based cross-subject emotion recognition using multi-source domain transfer learning. *Biomedical Signal Processing and Control*, 84:104741, 2023.

[23] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

[24] R. Sharma, R. B. Pachori, and P. Sircar. Automated emotion recognition based on higher order statistics and deep learning algorithm. *Biomedical Signal Processing and Control*, 58:101867, 2020.

[25] Q. She, C. Zhang, F. Fang, Y. Ma, and Y. Zhang. Multisource associate domain adaptation for cross-subject and cross-session eeg emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[26] T. Song, W. Zheng, P. Song, and Z. Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

[27] M. Stikic, R. R. Johnson, V. Tan, and C. Berka. Eeg-based classification of positive and negative affective states. *Brain-Computer Interfaces*, 1(2):99–112, 2014.

[28] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen. Eeg-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing*, 2020.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Y. Wang, S. Qiu, D. Li, C. Du, B.-L. Lu, and H. He. Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition. *IEEE/CAA Journal of Automatica Sinica*, 9(9):1612–1626, 2022.

[31] Z.-M. Wang, S.-Y. Hu, and H. Song. Channel selection method for eeg emotion recognition using normalized mutual information. *IEEE Access*, 7:143303–143311, 2019.

[32] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen. Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[33] Z. Yang, A. Kay, Y. Li, W. Cross, and J. Luo. Pose-based body language recognition for emotion and psychiatric symptom interpretation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 294–301. IEEE, 2021.

[34] H. J. Yoon and S. Y. Chung. Eeg-based emotion estimation using bayesian weighted-log-posterior function and perceptron convergence algorithm. *Computers in biology and medicine*, 43(12):2230–2237, 2013.

[35] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[36] Z. Zhang, S.-h. Zhong, and Y. Liu. Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*, 2022.

[37] F. Zheng, B. Hu, X. Zheng, C. Ji, J. Bian, and X. Yu. Dynamic differential entropy and brain connectivity features based eeg emotion recognition. *International Journal of Intelligent Systems*, 37(12):12511–12533, 2022.

[38] W.-L. Zheng and B.-L. Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.