Prior-Driven Dynamic Brain Networks for Multi-modal Emotion Recognition

Chuhang Zheng, Wei Shao, Daoqiang Zhang, and Qi Zhu(✉)

College of Computer Science and Technology, Key Laboratory of Brain-Machine
Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics
and Astronautics, Nanjing 211106, China
zhuqinuaa@163.com

Abstract. Emotions are closely related to many mental and cognitive diseases, such as depression, mania, Parkinson's Disease, etc, and the recognition of emotion plays an important role in diagnosis of these diseases, which is mostly limited to the patient's self-description. Because emotion is always unstable, the objective quantitative methods are urgently needed for more accurate recognition of emotion, which can help improve the diagnosis performance for emotion related brain disease. Existing studies have shown that EEG and facial expressions are highly correlated, and combining EEG with facial expressions can better depict emotion-related information. However, most of the existing multi-modal emotion recognition studies cannot combine multiple modalities properly, and ignore the temporal variability of channel connectivity in EEG. In this paper, we propose a spatial-temporal feature extraction framework for multi-modal emotion recognition by constructing prior-driven Dynamic Functional Connectivity Networks (DFCNs). First, we consider each electrode as a node to construct the original dynamic brain networks. Second, we calculate the correlation between EEG and facial expression through cross attention, as a prior knowledge of dynamic brain networks, and embedded to obtain the final DFCNs representation with prior knowledge. Then, we design a spatial-temporal feature extraction network by stacking multiple residual blocks based on 3D convolutions, and non-local attention is introduced to capture the global information at the temporal level. Finally, we adopt the features from fully connected layer for classification. Experimental results on the DEAP dataset demonstrate the effectiveness of the proposed method.

Keywords: Multi-modal emotion recognition · Dynamic brain networks · Cross attention · Non-local attention · 3D convolutions

1 Introduction

In healthcare, affective computing can help measure the psychological state of patients automatically, especially for those with cognitive deficits. For example, the emotional state of hospitalized patients contributes to the early diagnosis

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 389–398, 2023.
https://doi.org/10.1007/978-3-031-43993-3_38

of Parkinson's Disease (PD) [11]. In addition, for patients with neurological diseases, since neurological diseases are degenerative in nature, resulting in unstable cognitive function. The patients may not notice the symptoms of their disease, such as changes in their mood. Recent clinical diagnosis standards rely on the patients' self reports of their feelings to emotional disorders, but it may not be very accurate and stable. Therefore, we need to develop data-driven emotion identification method to improve the diagnosis of these disorders.

As EEG signals are directly related to high-level cognitive processes, EEG-based emotion recognition draws increasing attention in recent years [1]. Song et al. [15] proposed a dynamic graph convolutional network, which trained neural networks to dynamically learn the internal relationships between different EEG channels and extract more discriminative features. Zhang et al. [23] proposed a self-attention network to jointly model both local and global temporal information of EEG to reduce the effect of noise at the temporal level. These efforts do not take advantage of the complementary information between the modalities, which limits the performance of the model. Recently, a lot of works shown multi-modal data can provide complementary information to improve emotion recognition performance. Wang et al. [20] combined transformer encoders with attention based fusion to integrate EEG and eye movement data for emotion recognition. Ma et al. [10] designed a multi-modal residual long short-term memory network (MMResLSTM) to learn the correlation between EEG and peripheral physiological on multi-modal emotion recognition. However, the above work ignores correlations between EEG channels and fails to provide interpretable fusion model.

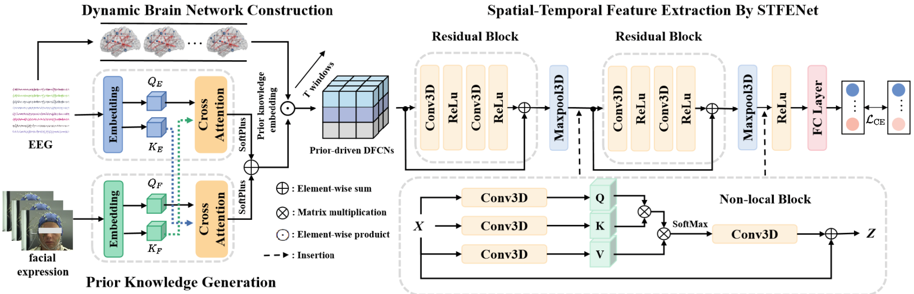

Fig. 1. Framework of our proposed multi-modal emotion recognition method.

Brain network analysis has been widely used in the field of disease diagnosis [8,22], which can describe the complex spatial relationships between brain regions of the brain. In recent years, researchers have migrated brain networks into emotion recognition. Wang et al. [21] implemented PageRank algorithm to rank the importance of brain network nodes, and screened important channels in emotion recognition according to the weight of channels. Huang et al. [6] proposed a novel neural decoding framework, which builds a bridge between emotions and brain regions, and captures their relationships by performing embedding propagation. However, the methods mentioned above regard the structure of brain network as static, ignoring that the variability of electrode channel connectivity over time. Since the multi-modal data is obtained from the synchronous stimulus in the same time period, this temporal level dynamic is particularly important in the multi-modal emotion recognition. In addition, integrating the heterogeneous data of EEG and facial expression also poses challenge to multi-modal emotion classification.

To overcome the above limitations, we design a spatial-temporal feature extraction framework based on prior-driven dynamic brain networks and apply it to emotion recognition. Specifically, we treat each electrode of EEG as a node of brain network, and then the dynamic functional connectivity networks (DFCNs) is constructed by Pearson correlation coefficient under non-overlapping time window. Besides, we calculate the correlation between EEG and facial expression across modal channels by cross attention mechanism, as the prior knowledge of DFCNs, and then embed it to above model obtain the final DFCNs representation. Finally, we implemented residual blocks and non-local attention to construct STFENet, so as to extract complex spatial-temporal feature and preserve the long-range dependencies in the time series.

2 Method

Figure 1 shows the framework of our approach, including four parts, i.e., the construction of dynamic brain networks, the representation and learning of cross-modal correlation between EEG and facial expression, the embedding of correlation into DFCNs as prior knowledge, and the extraction of spatial-temporal features of DFCNs for emotion recognition based on 3D convolutions.

Dynamic Brain Networks Construction: Functional Connectivity Networks (FCNs) ignore the temporal changes of brain connectivity. In this paper, we construct Dynamic Functional Connectivity Networks (DFCNs) to solve the above problem. First, each subject's EEG data can be represented as $X^E \in \mathbb{R}^{P \times T \times D}$, where P represents the number of channels, T is the number of time windows, and D represents the feature dimension. The t-th subsequence feature of P channels can be represented as a matrix $x(t) = [x_1(t), x_2(t), \cdots, x_p(t)] \in R^{P \times D}$, where $x_i(t) \in R^D$ represents the t-th subsequence feature extracted from the EEG time series of the i-th channel. According to the divided non-overlapping sliding time window, we build a functional connectivity network (i.e., matrix) by computing Pearson correlation coefficient between EEG from a pair of channels within the t-th time window:

$$C_{ij}(t) = \frac{\operatorname{cov}\left(x_i(t), x_j(t)\right)}{\sigma_{x_i(t)} \sigma_{x_j(t)}} \tag{1}$$

where cov denotes the covariance between two vectors, $\sigma_{x_i(t)}$ denotes the standard deviation of vector $x_i(t)$, $x_i(t)$ and $x_j(t)$ represent the EEG of a pair of channels i and j within the t-th time window, respectively. Thus, the original DFCNs of each subject $DFCNs_{\text{original}} = [C(1), C(2), \cdots, C(T)]^T \in R^{T \times P \times P}$ consists of T transient FCNs.

Prior Knowledge Embedding: Most of the existing multi-modal emotion recognition works aim to extract the features of different modalities respectively for fusion, which always lost the correlation between modalities. Existing studies have found there is high correlation between EEG and facial expression [5,12–14], but it is still challenging to find an appropriate way to fuse them. Therefore, we calculate the correlation of different modality data as prior knowledge to embed the previously constructed DFCN. Specifically, for each subject, $x^E \in R^{T \times P_E \times D}$, $x^F \in R^{T \times P_F \times D}$ represents the EEG and facial expression modality respectively, where T represents the number of time Windows, P_E and P_F represents the number of channels, and D represents the feature dimension. We perform different transformations by linear mapping pairs x^E and x^F:

$$Q_E = X^E W_Q^E, K_E = X^E W_K^E \tag{2}$$

$$Q_F = X^F W_Q^F, K_F = X^F W_K^F \tag{3}$$

where W_Q and W_K are the parameter matrices used to generate query and key, which are updated through network back-propagation during model training. We determined correlation scores across channel-dimension between modalities based on cross attention, by treating one modality as query and the other as key:

$$\text{Cor}(E, F) = \frac{Q_E K_F^T}{\sqrt{d_1}}, \text{Cor}(F, E) = \frac{Q_F K_E^T}{\sqrt{d_2}} \tag{4}$$

where, $\text{Cor}(E, F)$ and $\text{Cor}(F, E)$ represents the correlation score between the cross-modality channels, d_1, d_2 are normalized parameters equal to the dimension of K. It is worth noting that softmax is applied to the scoring weight of the equation Eq. 4. However, softmax proved to be overconfident in its results, which would result in the correlation scores of certain time windows being too high or too low, affecting the reliability of the prior knowledge. Therefore, we improve softmax to softplus to solve this problem while ensuring that the correlation matrix is non-negative. The calculated correlation matrix is as follows:

$$\text{Cor} = \text{softplus}(\text{Cor}(E, F) + \text{Cor}(F, E)) \tag{5}$$

At this point, we obtain the correlation between the cross-modal channels, and use it as the prior knowledge of DFCNs construction. We embed the modified prior knowledge into the previously constructed DFCNs by element-wise product:

$$DFCNs = DFCNs_{\text{original}} \odot \text{Cor} \tag{6}$$

where $\odot$ represents element-wise product. By the embedding of prior knowledge, we obtain the discriminative DFCNs representations with prior knowledge.

Spatial-Temporal Feature Extraction: Different from static brain networks, DFCNs can not only describe brain connectivity, but also contain the temporal-level volatility of brain connectivity. Most of the existing methods focus on extracting the temporal and spatial features of EEG separately, and concat them for feature fusion, which ignores the dynamic variations of electrode connectivity in the temporal dimension. 2D convolution has been widely used in the field of computer vision, but it is challenging to capture information at the temporal level. Previous studies has shown that 3D convolution operations can better model spatial information in continuous sequences [3,19]. So, we introduce 3D convolution to extract spatial-temporal feature of DFCNs simultaneously. Considering the DFCNs representation $X \in R^{C \times T \times P \times P}$ of each subject, where C is the number of channels, T represents the number of time windows, and P represents the number of electrode channels, then the m-th feature representation of the location (T, P, P) calculated by 3D convolution in space can be represented as:

$$v_m^{T,P,P} = \partial \left(b_m + \sum_{\sigma} \sum_{\varepsilon=0}^{P-1} \sum_{\rho=0}^{P-1} \sum_{\varphi=0}^{T-1} w_{c',m}^{T',P',P'} v_{c'}^{T+T',P+P',P+P'} \right) \tag{7}$$

where σ is the assigned activation function, b_m is the deviation, $w_{c',m}^{T',P',P'}$ represents the weight of the convolution kernel connected by the c'-th stacking channel to the feature representation of the position (T, P, P), and $v_{c'}^{T+T',P+P',P+P'}$ represents the characteristic value of the c'-th stacking channel at the position (T, P, P). To better capture the spatial-temporal topological structure in DFCNs, inspired by ResNet's remarkable success [4], we build a deeper network by stacking multiple residual blocks. A spatial-temporal feature extraction network (STFENet) is designed to extract spatial-temporal features of the DFCNs. The construction of STFENet is shown in the second half of Fig. 1. A residual block is used as the basic block, which includes two 3D convolutions, two activation functions and a residual connection. 3D Maxpooling is adopted between the multiple stacked residual block.

Since the operation of convolution will eventually focus on local areas, long-range dependencies which describe luxuriant emotion-related information will be lost to some extent. To solve this problem, we further introduce non-local block [18] to preserve information after the maxpooling layer. For a given input, non-local attention performs two different transformations:

$$\theta\left(x_i\right) = W_\theta x_i, \phi\left(x_i\right) = W_\phi x_i \tag{8}$$

where W_θ and W_ϕ is the weight to be learned, which is realized by 3D convolution in this paper. Then, non-local attention uses the self-attention term [17] to calculate the final features with the help of softmax:

$$y = \operatorname{softmax}\left(x^T W_\theta^T W_\phi x\right) g(x) \tag{9}$$

where g is implemented by $1 \times 1 \times 1$ convolution in this paper. Then, the non-local block can be defined as:

$$z = W_z y + x \tag{10}$$

where "$+x$" denotes the residual connection, and W_z represents the weight matrix. By the STFENet, we finally effectively extract the spatial-temporal emotion-related information in prior-driven DFCNs for the identification of emotions.

Table 1. Comparisons of different methods on DEAP dataset (Mean/Std%).

	method	Valence		Arousal	
		ACC	F1	ACC	F1
EEG based methods	SVM	58.88/11.47	67.81/13.28	58.05/15.26	66.71/20.54
	GSN	63.77/07.25	66.19/12.20	63.51/11.36	68.03/15.40
	DGCNN	63.28/08.15	65.17/10.22	62.61/12.93	69.22/15.17
Multi-modal based methods	MKL	59.02/12.94	67.37/16.97	58.96/13.89	64.78/18.69
	DCCA	61.04/11.73	65.46/12.29	59.58/11.89	65.78/16.98
	MMResLSTM	64.67/10.57	68.36/11.50	63.25/12.38	67.32/15.92
	ETF	64.63/09.35	66.35/13.41	65.64/10.29	66.63/16.97
	Ours	**67.36/05.58**	**69.17/09.01**	**68.47/08.45**	**74.68/13.36**

3 Experiment Results

Emotional Database: The DEAP dataset [9] collected EEG data from 32 healthy participants. The volunteers were asked to watch 40 one-minute videos and collect EEG signals from the subjects. The facial states of the first 22 subjects were recorded simultaneously. All participants rated each video on a 1–9 scale with the indicators, i.e. arousal, valence, dominance. We choose 18 subjects with both EEG signals and a complete facial video for the experiment. Same as many state-of-the-art studies [2], we turn the identification task into binary classification problem by setting the evaluation threshold of 5.

Data Pre-processing: For the EEG data, The 32-channel EEG signal with a duration of 63 s is down-sampled to 128HZ, and remove the first 3 s pre-trial baseline. Power spectral density (PSD) features is extracted from 3 s time windows with non-overlap through the Welch method in EEG, and 5 frequency bands are adopted, i.e. theta (4–8 Hz), slow alpha (8–10 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30+ Hz) [9]. For the facial video data, referring to [12], we utilize OPENFACE to extract expression features from facial videos, including 3 face positions relative to the camera, 3 head position, 6 eye gaze directions and 17 facial action units. Similar to EEG, the face sequences are divided according to the 3-second non-overlapping sliding time window and the average value of each feature is taken.

Experiment Settings: In our experiment, we adopt the leave one subject out (LOSO) cross-validation strategy to verify the effectiveness of our method. Specifically, the samples are divided into 18 non-overlapping parts according to the subjects. The samples of one subject are selected as the test set while the remaining subjects are selected as the training set for each cross-validation. This process is repeated 18 times, and the average performance of the cross-validation is taken as the final result. Identification performance is measured by accuracy (ACC) and F1-Score. The proposed method is based on the Pytorch implementation, and the model mentioned in this study is trained on a single GPU (NVIDIA GeForce RTX3080). Adam algorithm is used to optimize this method, and the learning rate and batch size are set to 0.001 and 40, respectively.

Results and Discussion: We evaluate the performance of our method by calculating ACC and F1 on both valence and arousal. We also compare our method with many comparison methods, which can be divided into two categories: EEG based methods and multi-modal based methods. More specifically, EEG based methods are Support vector machine (SVM), GraphSleepNet (GSN) [7], Dynamical Graph-CNN (DGCNN) [15]. Multi-modal based methods include Multi-kernel learning (MKL), Deep-CCA (DCCA), MMResLSTM [10]. Emotion transformer fusion (ETF) [20]. For quantitative results in Table 1, firstly, most of the multi-modal based methods achieve higher performance than EEG-based methods, which shows the advantage of complementary information from multiple modalities. Secondly, our proposed method achieve the best emotion recognition performance. On valence and arousal, the average ACC and F1 of our method reached 67.36%, 69.17% and 68.47% and 74.68% respectively. The main reason for the superiority of our method is that we can not only use multi-modal data as prior knowledge to guide the construction of DFCNs, but also extract discriminate spatial-temporal features.

Table 2. Ablation study of our proposed method (Mean/Std%).

method	Valence		Arousal	
	ACC	F1	ACC	F1
Baseline	61.11/11.35	64.81/13.40	60.91/12.45	65.10/18.99
w/o PKE	64.15/09.73	67.19/12.48	65.58/10.18	68.12/16.10
w/o NL block	66.12/08.60	66.92/10.36	66.24/09.28	72.31/14.63
w/o STFENet	65.51/08.19	67.18/11.29	67.10/10.15	70.83/15.84
Ours	**67.36/05.58**	**69.17/09.01**	**68.47/08.45**	**74.68/13.36**

To evaluate the effectiveness of the different modules of our framework, we conduct several ablation experiments on DEAP dataset. Our method mainly contains two modules, Prior knowledge embedding (PKE) module and STFENet. Besides, we also evaluate the contribution of non-local block in STFENet. As can be seen in Table 2, every module used in our framework greatly improve the performance compared with baseline model, with an increase of 6.25%, 4.36% and 7.56%, 9.58% for ACC and F1 on valence and arousal, respectively. It can be seen that both non-local block and STFENet demonstrate the better performance of our proposed method. The reason lies in that STFENet is able to extract complex spatial-temporal feature, and non-local block of STFENet helps it preserve the long-range dependencies in the time series. Moreover, when we remove PKE module from our method, there comes a performance degradation. It suggested that the prior knowledge has vital guiding significance for the construction of DFCNs, so that it can better express emotion-related information.

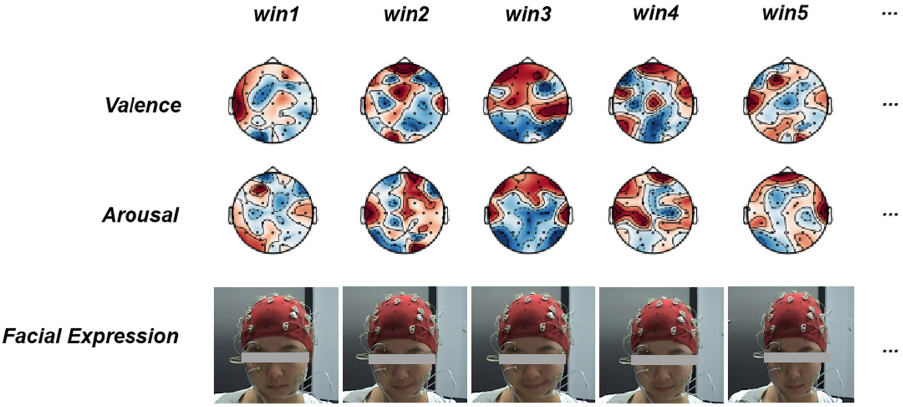

Fig. 2. Visualization of facial expression and its channel correlation with EEG in different time windows.

In addition, to further verify the feasibility of the prior knowledge embedded in our method, we visualize the facial expression of subject over several time windows and its channel correlation with EEG, as shown in Fig. 2. Firstly, from the channel correlation topographic map, the deeper the red of the brain area, the higher the correlation between EEG and facial expression. Conversely, the deeper the blue, the lower the correlation. On valence and arousal, high correlation areas focus on the binaural and prefrontal regions, which is in line with existing medical cognition [16]. As the stimulation method adopted by DEAP dataset is musical stimulation, the binaural region is highly activated. The prefrontal lobe plays a crucial role in emotional mobilization [2]. The experimental results show that our method can mine electrode channels that are highly correlated with emotion to provide prior knowledge to guide the construction of dynamic brain networks. Combined with the experimental results after removing PKE of our method in Table 2, it can be seen that embedding prior knowledge can achieve better emotion recognition performance. Therefore, the prior knowledge can better describe emotion-related information.

4 Conclusion

In this paper, we develop a spatial-temporal feature extraction framework based on prior-driven DFCNs for multi-modal emotion recognition. In our approach, not only the connectivity between EEG channels but also the dynamics of connectivity over time are jointly learned. Besides, we also calculate the correlation across modalities via cross attention to guide the construction of DFCNs. In addition, we build STFENet based on 3D convolution to model the spatial-temporal features contained in DFCNs to extract emotion-related spatial-temporal information and preserve the long-range dependencies in the time series. Experimental results show that our method outperforms the state-of-the-art methods.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Nos. 62076129, 62136004, and 62276130), the Key Research and Development Plan of Jiangsu Province (No. BE2022842), and the Fundamental Research Funds for the Central Universities (No. NS2023051).

References

1. Cai, Q., Cui, G.C., Wang, H.X.: EEG-based emotion recognition using multiple kernel learning. Mach. Intell. Res. **19**(5), 472–484 (2022)
2. Du, X., et al.: An efficient LSTM network for emotion recognition from multichannel EEG signals. IEEE Trans. Affect. Comput. **13**(3), 1528–1540 (2020)
3. Guo, S., Lin, Y., Li, S., Chen, Z., Wan, H.: Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting. IEEE Trans. Intell. Transp. Syst. **20**(10), 3913–3926 (2019)
4. He, F., Liu, T., Tao, D.: Why ResNet works? Residuals generalize. IEEE Trans. Neural Netw. Learn. Syst. **31**(12), 5349–5362 (2020)
5. Huang, X., et al.: Multi-modal emotion analysis from facial expressions and electroencephalogram. Comput. Vis. Image Underst. **147**, 114–124 (2016)
6. Huang, Z., Du, C., Wang, Y., He, H.: Graph emotion decoding from visually evoked neural responses. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13438, pp. 396–405. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_38
7. Jia, Z., et al.: GraphSleepNet: adaptive spatial-temporal graph convolutional networks for sleep stage classification. In: IJCAI, pp. 1324–1330 (2020)
8. Jie, B., Shen, D., Zhang, D.: Brain connectivity hyper-network for MCI classification. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 724–732. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10470-6_90
9. Koelstra, S., et al.: DEAP: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2011)
10. Ma, J., Tang, H., Zheng, W.L., Lu, B.L.: Emotion recognition using multimodal residual LSTM network. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 176–183 (2019)
11. Prell, T., et al.: Specialized staff for the care of people with Parkinson's disease in Germany: an overview. J. Clin. Med. **9**(8), 2581 (2020)

12. Rayatdoost, S., Rudrauf, D., Soleymani, M.: Multimodal gated information fusion for emotion recognition from EEG signals and facial behaviors. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 655–659 (2020)
13. Siddharth, S., Jung, T.P., Sejnowski, T.J.: Impact of affective multimedia content on the electroencephalogram and facial expressions. Sci. Rep. **9**(1), 16295 (2019)
14. Soleymani, M., Pantic, M., Pun, T.: Multimodal emotion recognition in response to videos. IEEE Trans. Affect. Comput. **3**(2), 211–223 (2011)
15. Song, T., Zheng, W., Song, P., Cui, Z.: EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Trans. Affect. Comput. **11**(3), 532–541 (2018)
16. Sun, Y., Ayaz, H., Akansu, A.N.: Multimodal affective state assessment using fNIRS+ EEG and spontaneous facial expression. Brain Sci. **10**(2), 85 (2020)
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
19. Wang, Y., et al.: 3d auto-context-based locality adaptive multi-modality GANs for pet synthesis. IEEE Trans. Med. Imaging **38**(6), 1328–1339 (2018)
20. Wang, Y., Jiang, W.B., Li, R., Lu, B.L.: 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), pp. 1575–1578. IEEE (2021)
21. Wang, Z.M., Zhang, J.W., He, Y., Zhang, J.: EEG emotion recognition using multichannel weighted multiscale permutation entropy. Appl. Intell. **52**(10), 12064–12076 (2022)
22. Yang, J., Zhu, Q., Zhang, R., Huang, J., Zhang, D.: Unified brain network with functional and structural data. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12267, pp. 114–123. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_12
23. Zhang, Y., Liu, H., Zhang, D., Chen, X., Qin, T., Zheng, Q.: EEG-based emotion recognition with emotion localization via hierarchical self-attention. IEEE Trans. Affect. Comput. 1 (2022)