DTCA: Dual-Branch Transformer with Cross-Attention for EEG and Eye Movement Data Fusion

Xiaoshan Zhang, Enze Shi, Sigang Yu, and Shu Zhang(✉)

Center for Brain and Brain-Inspired Computing Research, School of Computer Science,
Northwestern Polytechnical University, Xi'an, China
shu.zhang@nwpu.edu.cn

Abstract. Integrating Electroencephalography (EEG) and eye movements (EM) provides a comprehensive understanding of brain dynamics. However, effectively capturing essential information from EEG and EM poses challenges. Previous studies have investigated aligning and identifying correlations between them, yet they have not fully utilized the deep dynamic relationship and complementary features inherent in EEG and EM data. To address this issue, we propose the **D**ual-Branch **T**ransformer with **C**ross-**A**ttention (DTCA) framework. It encodes EEG and EM data into a latent space, leveraging a multimodal fusion module to learn the facilitative information and dynamic relationships between EEG and EM data. Utilizing cross-attention with pooling computation, DTCA captures the complementary features and aggregates promoted information. Extensive experiments on multiple open datasets show that DTCA outperforms previous state-of-the-art methods: 99.15% on SEED, 99.65% on SEED-IV, and 86.05% on SEED-V datasets. We also visualize confusion matrices and features to demonstrate how DTCA works. Our findings demonstrate that (1) EEG and EM effectively distinguish changes in brain states during tasks such as watching videos. (2) Encoding EEG and EM into a latent space for fusion facilitates learning promoted information and dynamic correlation associated with brain states. (3) DTCA efficiently fuses EEG and EM data to leverage their synergistic effects in understanding the brain's dynamic processes and classifying brain states.

Keywords: EEG, Eye Movement, Multimodal Fusion, Cross Attention, Brain Function Dynamics

1 Introduction

In myriad cognitive and emotional processes, the human brain generates diverse physiological signals linked to different brain states [1]. Electroencephalography (EEG) records internal brain signals with high temporal resolution, tending to reflect brain states. Eye movement (EM), governed by the cerebellum as brain states change, reflects external signals responding to physiological behaviors like rapid blinking and pupil dilation [2, 3]. Thus, the combination of EEG and EM harnesses their complementary strengths, providing a comprehensive understanding of the brain's dynamic. Technological advancements have paved the way for high-resolution EEG and eye-tracking devices, facilitating the simultaneous recording of EEG and EM data. This integration has shown promising results in enhancing emotion recognition [4–6], cognitive load assessment [1, 7], human-computer interaction [8, 9], and disease diagnosis, including Alzheimer's disease [10] and autism evaluation [11, 12]. Previous studies on EEG and EM fusion can be broadly categorized into two methods: feature splicing-based fusion and model-based fusion. For instance, Guo et al. [13] analyzed manually created features and then combined them for customer preference prediction. This splicing method ignores the complementary features between unimodal data. Zheng et al. [4] fused EEG and EM using a Boltzmann machine for emotion recognition, while Wang et al. [14] introduced an attention mechanism to combine EEG and EM. The model-based methods, while capable of learning complementary features, do not utilize the deep dynamic relationships and promoted information between EEG and EM.

To explore these challenges, we propose the **D**ual-Branch **T**ransformer with **C**ross-**A**ttention (DTCA), a novel framework that aims to enhance multimodal fusion efficiency. DTCA is designed to learn the dynamic correlation and capture the facilitation information and complementary features between EEG and EM, thereby improving understanding of the brain's dynamic processes. Our main contributions are as follows:

- We propose a novel multimodal fusion framework, which encodes EEG and EM data into a latent space, then utilizes an efficient multimodal fusion module to learn deep dynamic relationships from each modality and fuse them, where the fused features can be adapted to various brain states classification tasks.
- We propose DTCA, which utilizes the dual-branch transformer and incorporates cross-attention with pooling computation to facilitate the exchange of deep dynamic information and complementary features fusion between multimodalities.
- Extensive experimental results demonstrate the effectiveness, generalization, and superiority of our DTCA model on multiple open multimodal datasets.

2 Methods

2.1 Overview

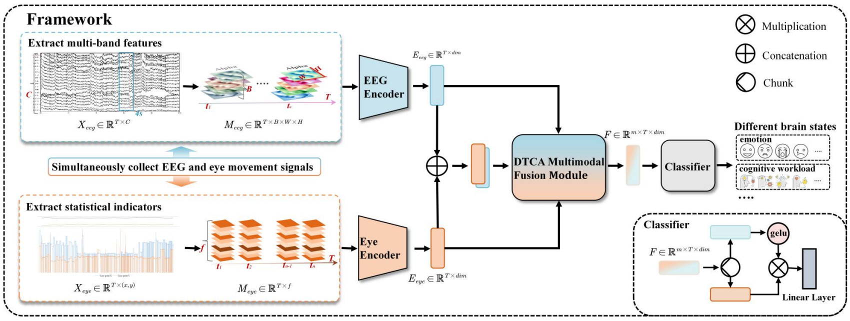

Framework
Extract multi-band features
C
4s
$X_{eeg} \in \mathbb{R}^{T \times C}$
Alpha
B
W
H
T
t_1
t_n
$M_{eeg} \in \mathbb{R}^{T \times B \times W \times H}$
Simultaneously collect EEG and eye movement signals
Extract statistical indicators
$X_{eye} \in \mathbb{R}^{T \times (x,y)}$
f
t_1
t_2
t_{n-1}
t_n
T
$M_{eye} \in \mathbb{R}^{T \times f}$
EEG Encoder
Eye Encoder
$E_{eeg} \in \mathbb{R}^{T \times dim}$
$E_{eye} \in \mathbb{R}^{T \times dim}$
DTCA Multimodal Fusion Module
$F \in \mathbb{R}^{m \times T \times dim}$
Classifier
Different brain states
emotion
cognitive workload
....
Multiplication
Concatenation
Chunk
Classifier
$F \in \mathbb{R}^{m \times T \times dim}$
gelu
Linear Layer

Fig. 1. Our multimodal fusion framework consists of four parts: an EEG encoder, an Eye encoder, a DTCA multimodal fusion module, and a classifier.

Our proposed framework is shown in **Fig. 1**. Given $X_{eeg} \in \mathbb{R}^{T \times C}$ and $X_{eye} \in \mathbb{R}^{T \times (x,y)}$, representing raw EEG with C channels and EM with coordinates (x,y) over the time window T. Then we extract f statistical indicators of EM $M_{eye} \in \mathbb{R}^{T \times f}$, and EEG feature maps $M_{eeg} \in \mathbb{R}^{T \times B \times W \times H}$ for B frequency bands with resolution (W,H). By feeding M_{eeg} and M_{eye} into the encoders of EEG and EM, we get the latent embeddings $E_{eeg} \in \mathbb{R}^{T \times dim}$ and $E_{eye} \in \mathbb{R}^{T \times dim}$ of dimension dim. The encoded unimodal latent embeddings are then fed into the DTCA multimodal fusion module, where they are mutually reinforced and fused. The fused embedding $F \in \mathbb{R}^{m \times T \times dim}$ with m modalities can be utilized in subsequent brain states classification tasks, such as emotion classification.

2.2 Unimodal Features Extraction

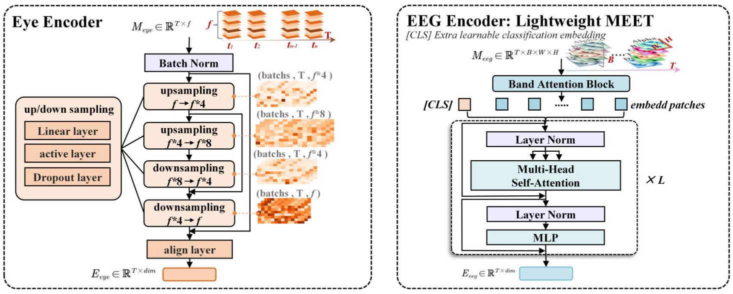


Fig. 2. Unimodal features encoders, left is the Eye encoder, right is the EEG encoder.

EEG encoder. For EEG feature extraction, we use the MEET [15] framework (on the right side of **Fig. 2**). Differential entropy (DE) features [16] are extracted from EEG signals in five frequency bands (delta: 1-4 Hz, theta: 4-8 Hz, alpha: 8-14 Hz, beta: 14-31 Hz, gamma: 31-50 Hz) using short-term Fourier transforms with a 4 s time window without overlapping. To handle electrode spatial distribution, we map 3D coordinates to a 2D plane using topology-preserve Azimuthal Equidistant Projection (AEP) [17]. Then, we transform the DE features vector into a $W \times H$ feature map through Clough-Tocher interpolation [18]. Consequently, we obtain EEG feature maps sized (T, 5, W, H). To temporally align with EM signals, we set T=1, rendering the temporal self-attention module in the original MEET obsolete. Hence, in our EEG encoder, only the spatial self-attention module is retained.

Eye encoder. A review by Skaramagkas et al. [19] suggested that EM indicators (e.g., pupil diameter, dispersion, fixation, blink, etc.) can provide valuable information for categorizing emotional and cognitive processes. Therefore, we use statistical indicators like pupil diameter, dispersion, fixation, and blink duration for EM features encoding. We design a specialized neural network, depicted on the left side of **Fig. 2** to analyze intricate connections among EM features. The eye encoder includes a batch normalization, two up/down sampling, and an aligned layer. Each sampling block has a linear layer, activation layer, and dropout layer. Upsampling enhances feature representation, while skip connections merge low and high-level features during downsampling.

2.3 Dual-Branch Transformer with Cross-Attention

In this section, we elucidate the DTCA methodology for feature fusion, as shown in **Fig. 3**. Traditional fusion techniques frequently neglect the deep dynamic relationships and complementary features between two modalities. Inspired by EMT [20], DTCA addresses this by leveraging transformers to capture global features and promote useful information exchange between modalities. It integrates interacted features using cross-attention and aggregates them with a pooling layer to enhance fusion.

Self-attention (SA) is the core component in Transformer. It allows the modeling of global dependencies in a sequence via scaled dot-product attention [21]. For the input sequence $E_t \in \mathbb{R}^{T_t \times d}$, we define the query as $Q_t = E_t W_Q$, key as $K_t = E_t W_K$, value as $V_t = E_t W_V$, where $W_Q, W_K \in \mathbb{R}^{d \times d_k}; W_V \in \mathbb{R}^{d \times d_v}$. Then SA can be formulated as follows:

$$SA(E_t) = softmax\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) V_t \tag{1}$$

Cross-attention (CA) involves two modalities, the query is from E_g the stack of E_{eeg} and E_{eye}, while the key and value are from the unimodal s, s presents E_{eeg} or E_{eye}, i.e., $Q_g = E_g W_Q$, $K_s = E_s W_K$, $V_s = E_s W_V$. In this way, CA bridges the information interaction between two modalities through cross-modal allocation of attention weights:

$$CA(E_g, E_s) = softmax\left(\frac{Q_e K_s^T}{\sqrt{d_k}}\right) V_s \tag{2}$$

Note that, for simplicity, we only present the formulation of single-head attention. In practice, we use multi-head SA or CA to allow the model to attend to information from different feature subspaces [21].

Finally, we briefly introduce the pooling layer for aggregating promoted information from different modalities to facilitate subsequent fusion. Specifically, we utilize an attention-based pooling layer to implement it:

$$F = softmax(v^T \tanh(W^T G_g^T + b)) G_g \tag{3}$$

where v, W, b are learnable parameters, G_g are stacked vectors from the dual-branch transformer, F is the final fused embedding. After obtaining the fused embedding, we utilize an active classifier, which integrates an activation layer, to enhance nonlinearity. Then, we update the parameters of the DTCA by computing the cross-entropy loss.

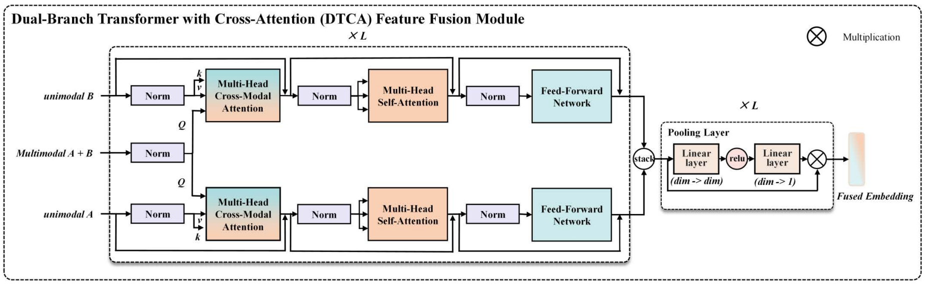


Fig. 3. Dual-Branch Transformer with Cross-Attention (DTCA) Feature Fusion Module

3 Experimental Results

3.1 Experimental Settings

To comprehensively evaluate our proposed model, we validate it on three multimodal datasets: SEED [22], SEED-IV [4], and SEED-V [6] obtained from Shanghai Jiaotong University's BCMI Lab under a data usage agreement. The SEED dataset includes fifteen carefully chosen Chinese film clips depicting happy, neutral, and sad emotions, with fifteen participants across three sessions. SEED-IV comprises 24 movie clips featuring various emotions, while SEED-V adds disgust emotion, with sixteen participants across three sessions. EEG signals (62-channel) and EM signals were recorded using the ESI Neuroscan system and SMI's eye-tracking glasses. For further details on the dataset, please refer to the official website of the SEED series datasets[1].

We follow the train/test protocols delineated in the original papers of each dataset, partitioning method is consistent with previous studies [4, 6, 22]. We use the first 9 (SEED) or 16 (SEED-IV) trials as the training data and the rest 6 (SEED) or 8 (SEED-IV) trials as the test data. Since subjects took part in the experiment for three sessions, we train each session separately. The results are reported as the average performance across all the subjects. For SEED-V, we conduct threefold cross-validation, concatenating features extracted from the first five clips across three sessions for each fold [5].

In our experiments, the time window for EEG feature maps and EM features is set to 1 s, meaning that each one-second segment of EEG feature maps and EM features is one sample. The eye encoder includes a Gelu activation function and a dropout rate of 0.2. For the EEG encoder, using the pre-processed brain images with 32×32 resolution as input. Unlike the MEET's configuration, we use a lightweight EEG Transformer with a depth of 1 and 2 heads. Each transformer in a DTCA with a depth of 2 and 4 heads. The batch size is set to 128. For training, we use the SGD optimizer with momentum 0.9 and weight decay 0.00005, along with the cosine descent algorithm for learning rate updates. We maintain a consistent random seed value of "123" for both training and testing. All experiments are conducted using PyTorch 1.8.0 on an NVIDIA RTX3090 with CUDA 11.1.

3.2 Effectiveness Evaluation of DTCA

Using the setup from Section 3.1, we conduct experiments on SEED, SEED-IV, and SEED-V datasets to compare the performance of EEG-only, EM-only, and multimodal. In the unimodal experiments, we only use either the EEG encoder or the EM encoder, without employing the cross-attention mechanism. We also compare our proposed DTCA with other methods in **Table 1**.

In **Table 1**, the bold formatting means the highest Accuracy (ACC) and Standard Deviation (STD) across each dataset, highlighting that our method achieves the top ACC on all datasets.

[1] https://bcmi.sjtu.edu.cn/~seed/index.html

Among unimodal methods, Xu's Single Eye [23] used only eye movements for classification, showing their potential for emotion differentiation. PR-PL [24], BiHDM [25], and MEET specialize in innovative EEG feature extraction. Among them, MEET is the current state-of-the-art method on SEED and SEED-IV datasets. The comparison results are shown in **Table 1**, our method outperforms MEET, by 0.39% and 5.57%. Additionally, it exhibits significant improvements over Single Eye, with increases of 18.13%, 23.91%, and 12.39% on SEED, SEED-IV, and SEED-V, respectively.

Table 1. Comparison of average accuracy and standard deviations (%) of each single modality and multimodality on different datasets

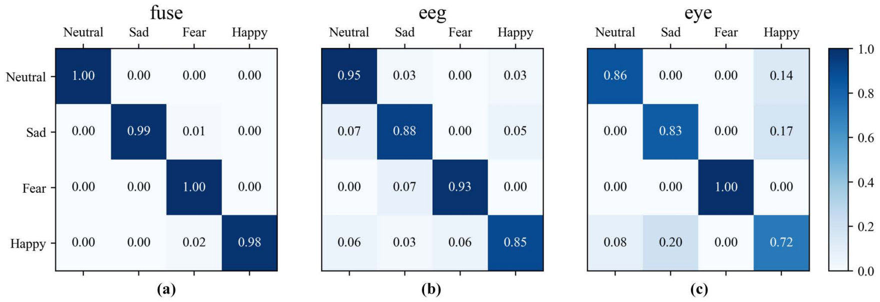
Method	Modality	SEED	SEED-IV	SEED-V
		(ACC ± STD)	(ACC ± STD)	(ACC ± STD)
Single Eye [23]	EM	81.02 ± 8.04	75.74 ± 6.66	73.66 ± 6.05
BDAE [4, 5]	EEG and EM	91.0 ± 8.9	85.1 ± 11.8	79.70 ± 4.76
BiHDM [25]	EEG	93.12 ± 6.06	74.35 ± 14.09	-
PR-PL [24]	EEG	94.84 ± 9.16	83.33 ± 10.61	-
DCCA [6]	EEG and EM	94.6 ± 6.2	87.5 ± 9.2	85.3 ± **5.6**
DCCA-FCP [26]	EEG and EM	95.08 ± 6.42	-	84.51 ± 5.11
MEET [15]	EEG	98.76 ± **0.78**	94. 08 ± 2.33	-
Ours	EM	84.33 ± 12.4	89.71 ± 7.3	71.07 ± 7.9
	EEG	96.68 ± 6.5	94.19 ± 3.8	71. 71 ± 8.7
	EEG and EM	**99.15** ± 2.9	**99.64** ± **0.5**	**86.05** ± 6.1

Among multimodal methods, BDAE and DCCA are currently the most widely used for fusing EEG and EM features. Zheng et al. [4], introduced the bidirectional deep self-encoder (BDAE) to fuse EEG and EM, while DCCA [6] aims to maximize the correlation between the two modalities by jointly learning their nonlinearity. Compared to these methods (refer to **Table 1** for results), DTCA achieves improvements of 4.07%-8.15% on SEED, 12.15%-14.55% on SEED-IV, and 0.75%-6.35% on SEED-V. In summary, DTCA demonstrates superior accuracy and stability compared to both unimodal and multimodal methods, and it yields the best results across all three datasets, indicating strong generalization ability. The above results also demonstrate that integrating EEG and EM enables accurate and robust discrimination of changes in brain states during emotional video viewing.

Table 1 reveals an interesting trend: while most multimodal methods perform slightly worse on SEED than EEG-only methods, as the task complexity rises, the multimodal methods notably surpass the unimodal one. For instance, DCCA is 0.24% lower than PR-PL on SEED, but 4.17% higher on SEED-IV. This indicates that incorporating multiple modalities can effectively handle more intricate and varied tasks.

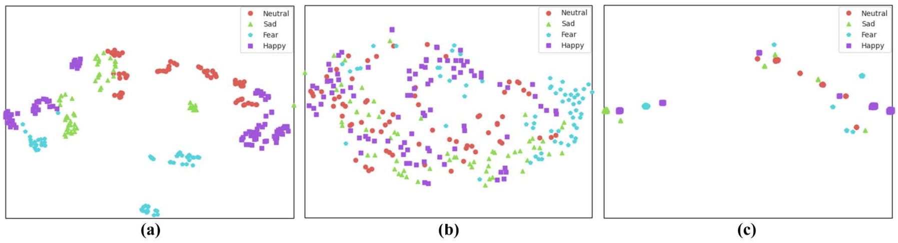
To analyze the complementary features between EEG and EM data, we generate confusion matrices for both unimodal and multimodal. In **Fig. 4**., the horizontal axis represents true labels, while the vertical axis shows predicted labels. Unimodal analysis reveals confusion between similar emotions, such as "Happy" being misclassified as "Neutral". EEG is effective in discerning "Neutral", while EM excels in categorizing "Fear", possibly due to noticeable changes in pupil dilation during fear, and EEG exhibits smoother patterns in a neutral state. By fusing EEG and EM through DTCA, their complementary features can be leveraged effectively. This fusion notably reduces confusion, fully capturing their complementary features and dynamic relationships.

Additionally, we visualize EEG, EM, and fusion features in **Fig. 5** to show the effectiveness of combining them. After reducing them to 2 dimensions via t-SNE, we observe that individual EEG and EM features appear fragmented and overlapping. In contrast, different categories of fused features show clearer boundaries, indicating a strong discriminative capability. This underscores DTCA's effectiveness in capturing deep dynamic relationships and extracting complementary features between modalities.



(a) (b) (c)

Fig. 4. Comparison of the confusion matrices of different modalities on the SEED-IV datasets. (a) multimodal, (b) EEG unimodal, (c) eye movements unimodal



(a) (b) (c)

Fig. 5. Visualization of different modal features. (a) represents fused multimodal features, (b) represents EEG features, (c) represents EM features

3.3 Ablation study

To verify the effectiveness of each module of DTCA, we conduct comprehensive ablation experiments. As mentioned in Section 2.3, we add cross-attention for feature fusion between modalities. The pooling layer aggregates information from different modalities with adaptable attention parameters, while the active classifier boosts classification accuracy with an activation function. We design ablation experiments for these three modules, all conduct on the SEED-IV dataset, and the results are shown in **Table 2**. Under the same experimental conditions, the inclusion of all three modules resulted in performance enhancement, with the pooling layer showing the most significant improvement, manifesting an increase of 2.04%.

In addition, to validate that DTCA outperforms other feature fusion methods, we compare several fusion approaches. Early fusion aligns and concatenates data from two modalities before inputting them into the model, utilizing the transformer. Middle fusion involves inputting two modalities into their respective encoders, followed by feature fusion through an attention module. Late fusion integrates the two modalities at the final classification stage, like integrated learning. From the results in **Fig. 6**, it is evident that DTCA significantly outperforms other fusion methods. This superiority can be attributed to DTCA's feature fusion interactions at early, intermediate, and late stages. In the early stage, the cross-attention mechanism assigns attention weights based on modality relevance, followed by calculating intra-modal attention weights through self-attention. Finally, the information from different modalities is aggregated in the pooling layer.

Table 2. Ablation study based on SEED-IV.

Pooling layer	Cross-attention	Active classifier	Acc(%)
✘	✘	✘	94.58
✔	✘	✘	96.62
✘	✔	✘	94.89
✘	✘	✔	94.72
✘	✔	✔	95.67
✔	✔	✘	96.97
✔	✘	✔	97.89
✔	✔	✔	**99.64**

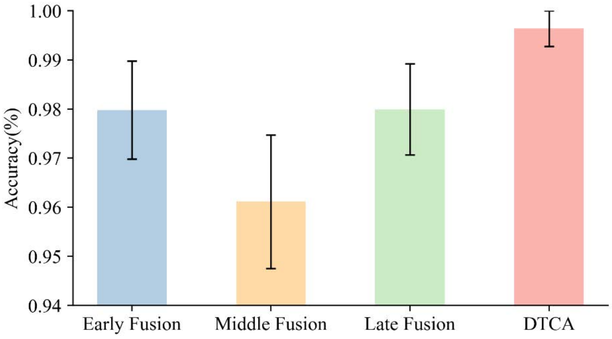


Fig. 6. Comparison results with different fusion methods based on SEED-IV

4 Conclusion

In this work, we demonstrate the effectiveness of incorporating EEG and EM data in understanding the brain's dynamic processes and accurately classifying brain states. Our proposed DTCA framework illustrates encoding EEG and EM into a latent space for fusion facilitates learning promoted information and dynamic correlation associated with brain states. The inclusion of cross-attention in DTCA, along with pooling computation, enhances the fusion of EEG and EM data, resulting in a significant improvement in brain states classification. Experimental results on three open multimodal datasets show that DTCA achieves state-of-the-art results of 99.15% on SEED, 99.65% on SEED-IV, and 86.05% on SEED-V datasets. One limitation of DTCA is that its transformer-based architecture necessitates sufficient training data to learn the dynamic relationships and fusion features between modalities. With the increasing convenience of acquiring EEG and EM data, we will expand our research by building larger datasets encompassing diverse brain states classification tasks. This will enable us to further validate the effectiveness of our proposed DTCA framework.

Acknowledgments. This work was supported by Foundational Research Project in Specialized Discipline (Grant No. G2024WD0146); First-Class Discipline Construction Project (Grant No. 23GH0104012); Faculty Construction Project (Grant No. 24GH0201148); National Natural Science Foundation of China (62376219 and 62006194).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Langer, N., Ho, E.J., Alexander, L.M., Xu, H.Y., Jozanovic, R.K., Henin, S., Petroni, A., Cohen, S., Marcelle, E.T., Parra, L.C., Milham, M.P., Kelly, S.P.: A resource for assessing information processing in the developing brain using EEG and eye tracking. Sci Data. 4, 170040 (2017).
2. Lin, C., Zhang, C., Xu, J., Liu, R., Leng, Y., Fu, C.: Neural correlation of EEG and eye movement in natural grasping intention estimation. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 31, 4329–4337 (2023).
3. Eckmiller, R.: Neural control of pursuit eye movements. Physiological Reviews. 67, 797–857 (1987).
4. Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., Cichocki, A.: EmotionMeter: A multimodal framework for recognizing human emotions. IEEE Trans. Cybern. 49, 1110–1122 (2019).
5. Zhao, L.-M., Li, R., Zheng, W.-L., Lu, B.-L.: Classification of five emotions from EEG and eye movement signals: Complementary representation properties. In: 2019 9th International IEEE/EMBS Conference on Neural Engineering. pp. 611–614. IEEE, San Francisco, CA, USA (2019).
6. Liu, W., Qiu, J.-L., Zheng, W.-L., Lu, B.-L.: Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Trans. Cogn. Dev. Syst. 14, 715–729 (2022).
7. Bodala, I.P., Kukreja, S., Li, J., Thakor, N.V., Al-Nashash, H.: Eye tracking and EEG synchronization to analyze microsaccades during a workload task. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 7994–7997 (2015).
8. Wang, Y., Yu, S., Ma, N., Wang, J., Hu, Z., Liu, Z., He, J.: Prediction of product design decision making: An investigation of eye movements and EEG features. Advanced Engineering Informatics. 45, 101095 (2020).
9. Zhu, S., Qi, J., Hu, J., Hao, S.: A new approach for product evaluation based on integration of EEG and eye-tracking. Advanced Engineering Informatics. 52, 101601 (2022).

10. Sriram, H., Conati, C., Field, T.: Classification of Alzheimer’s disease with deep learning on eye-tracking data. In: International conference on multimodal interaction. pp. 104–113. ACM, Paris France (2023).
11. Thapaliya, S., Jayarathna, S., Jaime, M.: Evaluating the EEG and eye movements for autism spectrum disorder. In: 2018 IEEE International Conference on Big Data. pp. 2328–2336 (2018).
12. Keles, U., Kliemann, D., Byrge, L., Saarimäki, H., Paul, L.K., Kennedy, D.P., Adolphs, R.: Atypical gaze patterns in autistic adults are heterogeneous across but reliable within individuals. Molecular Autism. 13, 39 (2022).
13. Guo, F., Li, M., Hu, M., Li, F., Lin, B.: Distinguishing and quantifying the visual aesthetics of a product: An integrated approach of eye-tracking and EEG. International Journal of Industrial Ergonomics. 71, 47–56 (2019).
14. 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. pp. 1575–1578. IEEE, Houston, TX, USA (2021).
15. Shi, E., Yu, S., Kang, Y., Wu, J., Zhao, L., Zhu, D., Lv, J., Liu, T., Hu, X., Zhang, S.: MEET: A multi-band eeg transformer for brain states decoding. IEEE Trans. Biomed. Eng. 1–12 (2023).
16. Duan, R.-N., Zhu, J.-Y., Lu, B.-L.: Differential entropy feature for EEG-based emotion classification. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering. pp. 81–84. IEEE, San Diego, CA, USA (2013).
17. Snyder, J.P.: Map projections: A working manual. U.S. Government Printing Office (1987).
18. Alfeld, P.: A trivariate clough—tocher scheme for tetrahedral data. Computer Aided Geometric Design. 1, 169–181 (1984).
19. Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N., Tripoliti, E., Marias, K., Fotiadis, D.I., Tsiknakis, M.: Review of eye tracking metrics involved in emotional and cognitive processes. IEEE Rev. Biomed. Eng. 16, 260–277 (2023).
20. Sun, L., Lian, Z., Liu, B., Tao, J.: Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Trans. Affective Comput. 1–17 (2023).
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2017).
22. Wei-Long Zheng, Bao-Liang Lu: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Mental Dev. 7, 162–175 (2015).
23. Yan, X., Zhao, L.-M., Lu, B.-L.: Simplifying multimodal emotion recognition with single eye movement modality. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1057–1063. Association for Computing Machinery, New York, NY, USA (2021).
24. Zhou, R., Zhang, Z., Fu, H., Zhang, L., Li, L., Huang, G., Dong, Y., Li, F., Yang, X., Liang, Z.: PR-PL: A novel transfer learning framework with prototypical representation based pairwise learning for EEG-based emotion recognition, (2022).
25. Li, Y., Wang, L., Zheng, W., Zong, Y., Qi, L., Cui, Z., Zhang, T., Song, T.: A novel bi-hemispheric discrepancy model for eeg emotion recognition. IEEE Trans. Cogn. Dev. Syst. 13, 354–367 (2021).

26. Wu, X., Zheng, W.-L., Li, Z., Lu, B.-L.: Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. J. Neural Eng. 19, 016012 (2022).