Heliyon 10 (2024) e34245

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

STDCformer: Spatial-temporal dual-path cross-attention model for fMRI-based autism spectrum disorder identification

Haifeng Zhang [a,b], Chonghui Song [a,*], Xiaolong Zhao [a], Fei Wang [c], Yunlong Qiu [a], Hao Li [a], Hongyi Guo [a]

[a] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[b] Division of Psychology, Nanyang Technological University, Singapore S639798, Singapore
[c] Department of Psychiatry, Affiliated Brain Hospital of Nanjing Medical University, Nanjing 210029, China

ABSTRACT

Resting-state functional magnetic resonance imaging (rs-fMRI) is a non-invasive neuroimaging technique widely utilized in the research of Autism Spectrum Disorder (ASD), providing preliminary insights into the potential biological mechanisms underlying ASD. Deep learning techniques have demonstrated significant potential in the analysis of rs-fMRI. However, accurately distinguishing between healthy control group and ASD has been a longstanding challenge. In this regard, this work proposes a model featuring a dual-path cross-attention framework for spatial and temporal patterns, named STDCformer, aiming to enhance the accuracy of ASD identification. STDCformer can preserve both temporal-specific patterns and spatial-specific patterns while explicitly interacting spatiotemporal information in depth. The embedding layer of the STDCformer embeds temporal and spatial patterns in dual paths. For the temporal path, we introduce a perturbation positional encoding to improve the issue of signal misalignment caused by individual differences. For the spatial path, we propose a correlation metric based on Gramian angular field similarity to establish a more specific whole-brain functional network. Subsequently, we interleave the query and key vectors of dual paths to interact spatial and temporal information. We further propose integrating the dual-path attention into a tensor that retains spatiotemporal dimensions and utilizing 2D convolution for feed-forward processing. Our attention layer allows the model to represent spatiotemporal correlations of signals at multiple scales to alleviate issues of information distortion and loss. Our STDCformer demonstrates competitive results compared to state-of-the-art methods on the ABIDE dataset. Additionally, we conducted interpretative analyses of the model to preliminarily discuss the potential physiological mechanisms of ASD. This work once again demonstrates the potential of deep learning technology in identifying ASD and developing neuroimaging biomarkers for ASD.

1. Introduction

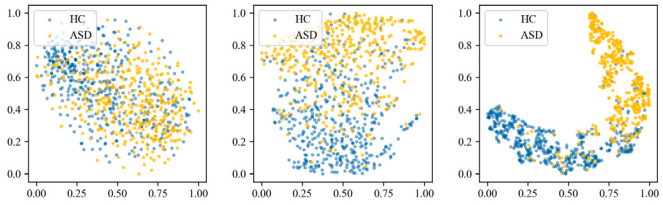
Resting-state functional magnetic resonance imaging (rs-fMRI) is a non-invasive and non-radioactive technique that reflects the neural activity of the brain. The potential of deep learning in analyzing the blood oxygen level-dependent (BOLD) time series of fMRI has been recognized [1]. The deep learning model holds promise in capturing specific dynamic features or abnormalities of brain-wide functional connectivity associated with neurodevelopmental disorders such as Autism Spectrum Disorder (ASD) from fMRI data [2].

The simultaneous emergence of studies emphasizing the temporal dynamics of fMRI signals (e.g., [3][4][5]) and studies highlighting the spatial correlation of fMRI signals (e.g., [6][7][8]) implies that both temporal and spatial features are crucial for identifying ASD. Consequently, there is a growing research interest that considers the joint representation of temporal and spatial patterns. Fig. 1 briefly shows the necessity of spatiotemporal information for identifying ASD. Fig. 1(a) and (b) illustrate that considering only temporal features [9] or only spatial features [10] results in fuzzy boundaries between ASD and HC representations. Fig. 1(c) demonstrates

* Corresponding author.
E-mail address: songchonghui@mail.neu.edu.cn (C. Song).

https://doi.org/10.1016/j.heliyon.2024.e34245
Received 1 April 2024; Received in revised form 5 June 2024; Accepted 5 July 2024

Available online 10 July 2024



(a) Transformer (b) GAT (c) STDCformer

Fig. 1. t-SNE plot of features fed into the classifier. (a) and (b) showcase results from the temporal model Transformer and the spatial model Graph Attention Network (GAT), respectively. (c) showcase the results from our spatial-temporal model STDCformer.

that our STDCformer, which integrates spatiotemporal information, enables more distinctly discrimination between ASD and healthy control group (HC).

Studies that simultaneously consider spatial and temporal patterns have also highlighted the importance of synergistic spatiotemporal information for fMRI-based ASD identification. However, capturing the spatiotemporal information upon which ASD relies from complex fMRI data poses challenges. Some studies have proposed frameworks that tightly couple temporal feature extraction functions with spatial feature extraction functions to learn the spatiotemporal patterns of fMRI [11][12][13]. These models deeply integrate spatiotemporal information. However, models that mix spatiotemporal information for analysis struggle to extract effective domain-specific representations and increase the difficulty of hyperparameter tuning. Therefore, some studies have proposed loosely coupled frameworks that separately capture temporal and spatial features on different branches to capture the spatiotemporal information of fMRI [14][15]. These studies have enhanced the ability to extract features capable of identifying ASD from intertwined spatiotemporal information by specially designing networks for temporal and spatial paths. However, these models do not emphasize the deep interaction between temporal and spatial features. Taken together, we reconsider whether we can simultaneously integrate the advantages of tightly coupled and loosely coupled models. Concurrently, we are also considering the ability to adapt to individualized noise at the temporal scale and to construct stable brain networks at the spatial scale.

To this end, we propose a dual-path cross-attention model, named STDCformer, which can interact explicitly deeply with spatiotemporal relationships while preserving pattern-specific information. Roughly, STDCformer initially embeds temporal and spatial features separately in dual paths, and the spatiotemporal dual-path features interact through cross-attention, integrated into spatiotemporal attention tensors for forward processing using 2D convolutions. Considering individual differences, heterogeneity, noise, and other issues in fMRI signals, we have purposefully designed a series of structures to address these challenges. In summation, the main contributions of this paper are as follows.

- We propose a novel dual-path cross-attention framework called STDCformer for more accurate fMRI-based ASD identification. STDCformer explicitly deeply integrates temporal and spatial information while preserving domain-specific patterns in the multivariate time series.
- In the temporal path embedding layer of STDCformer, we propose perturbation position encoding injected into token features to adapt to alignment biases caused by individual differences.
- In the spatial path embedding layer of STDCformer, we propose constructing whole-brain functional connections by channel-wise structural similarity of signal Gramian angular field images to finely describe the global dynamic structural correlations of whole-brain neural activities.
- In the self-attention layer of STDCformer, we design a dual-path cross-attention mechanism and propose integrating dual-path attention into a preserving spatiotemporal dimensions tensor and using 2D convolution for feedforward to reduce information loss and distortion.
- We extracted and analyzed fMRI-derived ASD neuroimaging biomarkers from our STDCformer.

2. Related work

2.1. Spatiotemporal feature extraction

In the context of existing fMRI-based ASD recognition, we categorize spatiotemporal deep learning into two types: tightly coupled spatiotemporal patterns and loosely coupled spatiotemporal patterns.

Tightly coupled spatiotemporal patterns refer to a paradigm where temporal and spatial information is interwoven. For instance, Azevedo et al. [11] and Park et al. [16] concatenate time feature extraction modules and spatial feature extraction modules. However, this approach may increase the difficulty of hyperparameter tuning and could lead to error accumulation. Therefore, some studies attempt to supplement spatial information in the temporal feature extraction model or transform the spatial feature extraction model into a dynamic form. Zhang et al. [17] designed a kernelized attention mechanism under the Transformer framework to jointly learn functional connectivity information and dynamic information of the brain. Deng et al. [12] introduced a spatial attention layer connected to a temporal attention layer under the Transformer framework to learn a spatiotemporal common representation. Gadgil et al. [13] considered the spatiality and dynamics of fMRI signals in the context of spatiotemporal graph convolutional networks. Xing

et al. [18] considered the spatiality and dynamics of fMRI signals in the context of dynamic graph convolutional networks. These models deeply integrate temporal and spatial information. Models where temporal and spatial patterns intertwine emphasize the potential synergistic mechanisms of spatiotemporal information. However, the complex entanglement of temporal and spatial patterns exacerbates the robustness challenges in analyzing fMRI data with heterogeneity and individualized biases. Additionally, the flexibility and scalability of such models also have limitations. Therefore, architectures with loosely coupled spatiotemporal patterns are favored.

Loosely coupled spatiotemporal modeling refers to a paradigm where time information and spatial information are modeled separately on different branches. For instance, Liu et al. [14] developed a spatiotemporal cooperative attention learning model with parallel temporal attention and spatial attention to model the association between fMRI and neurological disorders. Cui et al. [19] designed a dual-branch graph neural network to extract temporal and spatial features separately. Liu et al. [15] combined features extracted separately by dynamic graph convolution and LSTM to learn the spatiotemporal representation of time series. These methods emphasize the complementarity of time and space information by integrating heterogeneous information from different paths for competitive ASD recognition performance. Modeling time patterns and spatial patterns on different paths allows researchers to flexibility design architectures for corresponding branches tailored to the temporal and spatial characteristics. However, the loosely coupled model may not delve deeply into the interaction of spatiotemporal information compared to the tightly coupled model. The specific spatiotemporal patterns of highly heterogeneous ASD are unstable. The profound interaction of spatiotemporal features aids in addressing the challenge of ASD heterogeneity by modeling the evolution patterns of multiple information scales. Therefore, one of the key considerations in this work is how to deeply and explicitly interact with time-specific and space-specific information while preserving temporal and spatial specificity.

This work is based on a highly scalable Transformer architecture because we have found that models emphasizing temporal dynamics often outperform those emphasizing spatial relationships. We adapt the Transformer into dual paths to model temporal and spatial patterns. The dual-path embedding layers include modifications for adapting to the temporal dynamics and establishing whole-brain functional connections to emphasize temporal-specific and spatial-specific information. The dual-path attention layers include our designed explicit cross-attention and dual-path feature integration modules to emphasize deep interactions of spatiotemporal information.

2.2. Brain functional connectivity construction

The functional connectivity (FC) patterns represented by rs-fMRI data contain sensitive information for ASD identification. The graphs naturally adapt to brain networks, where nodes typically represent regions of interest (ROIs) in the brain, and edges represent functional associations between ROIs [7].

Quantifying FC between brain regions through Pearson correlation coefficients among ROIs quantifies the level of synchronization or coordination of activity across different ROIs [20]. However, some methods for measuring static functional connectivity (sFC) across the entire brain assume that brain activity remains constant over time [7][19]. This approach overlooks the dynamic changes in brain activity, capturing only overall average features. Therefore, dynamic FC (dFC) is considered a promising approach because it reflects the complex temporal characteristics of brain activity [15][21][22]. The evolving characteristics of brain networks over time captured by dFC include valuable sensitivity and specificity information for identifying brain anomalies. However, the effectiveness of dFC depends on the choice of the length of time windows and the sliding step size. Moreover, within shorter time windows, dFC may suffer from unstable connections due to noise, a concern heightened particularly for ASD, which exhibits higher heterogeneity. Although some studies have enhanced the stability of dFC through multiple templates [23] or multiple perspectives [8], this comes at the cost of increased computational complexity.

To address this, we propose a weakened form of dFC that considers the temporal dynamics of ROIs while maintaining the stability of sFC measurements. We thus consider the similarity between Gramian angular fields (GAFs) of ROIs. GAF combines phase and amplitude information of time series and is independent of sliding windows. GAF not only considers global features of signals but also captures subtle changes within signals. Additionally, the high resolution of GAF allows for the representation of richer nonlinear features within the signal.

3. Methods

This work introduces a dual-path cross-attention model named STDCformer for fMRI-based ASD identification, which comprehensively integrates spatial-temporal specific information and spatial-temporal depth interaction. The pipeline of STDCformer is illustrated in Fig. 2.

3.1. Problem statement

This work is treated as a multivariate time series classification task. We define X as a multivariate BOLD sequence with m time points for n regions of interest (ROIs), i.e., $X \in \mathbb{R}^{m \times n}$, and its corresponding label is y. Our primary objective is to comprehensively consider spatiotemporal representation to complete $X \mapsto y$.

3.2. Temporal feature embedding

In the temporal path feature embedding layer, we propose a method that injects perturbation position encoding (PPE) to address the signal alignment bias issue caused by individual differences significantly affecting the fMRI signal.

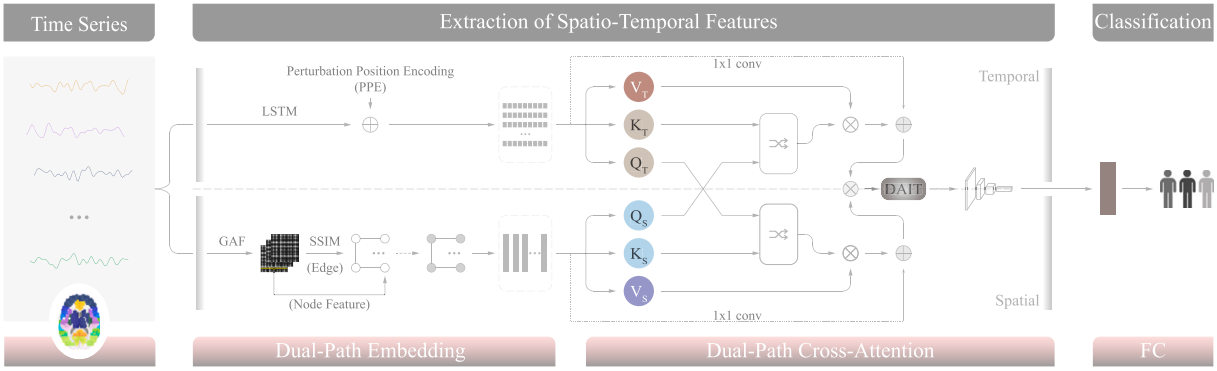


Fig. 2. Pipeline of the proposed STDCformer. The core of STDCformer lies in its dual-path embedding layers and dual-path cross-attention. In the temporal path embedding layer, perturbation positional encoding (PPE) is injected into token features extracted by LSTM. In the spatial path embedding layer following the graph convolutional paradigm, the graph topology is constructed based on GAF-SSIM. Inside the dual-path cross-attention block, query vectors from the temporal and spatial paths are exchanged, then the attentions of both paths are merged into a dual-path integrated tensor (DAIT) and forwarded using 2D convolution.

Due to variations in brain structure, brain function, head motion, and physiological noise among different individuals, fMRI signals may not be perfectly aligned in temporal or space, even in the case where preprocessing has been done using spatial normalization techniques. Therefore, there are limitations in the generalization of sequence position annotation across individuals using vanilla positional encoding. To address this issue, this work proposes an extremely simple and practical positional encoding method that reduces these limitations by adding a random subtle noise o_p to each position.

$$\mathrm{PPE} = \begin{cases} \sin\left(\dfrac{\mathrm{pos}+o_p}{10000^{\frac{i}{d_e}}}\right), \text{ if } i\%2 = 0 \\ \cos\left(\dfrac{\mathrm{pos}+o_p}{10000^{\frac{i}{d_e}}}\right), \text{ if } i\%2 = 1 \end{cases} \quad i \in \mathbb{Z}_{[0,m)}, \tag{1}$$

where pos is the position index in the input sequence, $\dim(o_p) = \dim(\mathrm{pos})$ and $\mathrm{PPE} \in \mathbb{R}^{m \times d_e}$. While appropriate $\dim(o_p)$ aids the model in accommodating individual noise, it should be noted that excessive random subtle noise may lead to distortion in positional encoding.

The introduction of position encoding with random subtle noise enables adaptation to small variations in the data, so it is more universally applicable for different individuals in terms of annotating sequence positions. This contributes to enhancing the model's adaptability and robustness to changes. Additionally, the random subtle noise introduces a level of randomness, so it aids the model in better capturing dynamic changes and uncertainties in the data. It is beneficial for handling dynamic changes in data, such as brain activity.

PPE is combined with token features just like other conventional methods. To compensate for the lack of global attention in capturing local details, we utilize LSTM to extract the dynamic features of the input's multi-dimensional time series. The embedding features of the temporal-path are

$$Z^T_{\mathrm{emb}} = \varepsilon_T(X) + \mathrm{PPE} \tag{2}$$

Here, $Z^T_{\mathrm{emb}} \in \mathbb{R}^{m \times d_e}$. ε_T is LSTM network.

3.3. Spatial feature embedding

The spatial feature embedding layer is a type of graph convolutional layer. Innovatively, we propose a signal spatial correlation measure based on the Structural Similarity Index between channel Gramian angular fields (GAF-SSIM) to construct whole-brain functional connectivity (BFC).

3.3.1. GAF-based whole-brain functional connectivity

We propose initially representing the time series of each channel as a Gramian Angular Field (GAF) image. In comparison to the temporal domain, GAF captures higher-order statistical information about the spatiotemporal structure of the signal while preserving the sequential information of the signal. Additionally, GAF is capable of processing signals with alignment offsets, which helps adapt to fMRI data with individual differences. GAF has two variants: GASF (Gramian Angular Summation Field) and GADF (Gramian Angular Difference Field) [24]. GADF emphasizes the short-term dynamic changes and fluctuations in time series data, while GASF focuses more on capturing the overall dynamic information and structure of time series data. Considering the multi-center bias of fMRI signals and the instability of local dynamic relationships in time series data caused by individual differences, we use GASF to describe the fMRI sequences for each brain region. For c-th region of X, note as X_c, its GASF is

$$\mathcal{X}_c = X_c^T \cdot X_c - \sqrt{I - \left(X_c\right)^2}^T \cdot \sqrt{I - \left(X_c\right)^2} \tag{3}$$

where I is the unit row vector.

Subsequently, we utilize the widely adopted Structural Similarity Index (SSIM) to assess the similarity of GAF images between channels as a proxy for the correlation between variables. SSIM takes into account the internal information of images and is more suitable for data with structural features such as GAF images. Moreover, SSIM is not dependent on absolute brightness and contrast levels of images and is insensitive to scaling and data noise. Thus, SSIM provides a more robust measure of similarity between GAF images across channels. The SSIM between c_1-th and c_2-th is

$$\Gamma_{c_1,c_2} = \frac{\left(2\mu_{c_1}\mu_{c_2} + \delta_1\right)\left(\sigma_{c_1c_2} + \delta_2\right)}{\left(\mu_{c_1}^2 + \mu_{c_2}^2 + \delta_1\right)\left(\sigma_{c_1}^2 + \sigma_{c_2}^2 + \delta_2\right)} \tag{4}$$

where μ_{c_1} and μ_{c_2} are the mean of $\mathcal{X}_{c_1}$ and $\mathcal{X}_{c_2}$, respectively. σ_{c_1} and σ_{c_1} are the standard deviation of $\mathcal{X}_{c_1}$ and $\mathcal{X}_{c_2}$, respectively. $\sigma_{c_1c_2}$ is the covariance of $\mathcal{X}_{c_1}$ and $\mathcal{X}_{c_2}$, δ_1 and δ_2 represent constants. Γ stands for Whole Brain Functional Connectivity (BFC).

In summation, in the construction of BFC, GAF-SSIM provides a more detailed measurement of the non-linear structure of signals and perceptual similarity compared to other methods that calculate correlation within the temporal domain.

3.3.2. Graph convolutional-based feature embedding

In the graph convolutional network paradigm, we represent spatial features to capture the spatial topological structure of complex brain networks. Graph convolutional networks can flexibly adapt to the irregular connectivity patterns, such as BFC, which facilitates the reflection of advanced abstract associations of neural activities between regions of interest (ROIs) in the brain. We obtain an adjacency matrix by thresholding the BFC, preserving only the essential information. Thresholding involves setting weak connections in BFC to zero, ensuring sparsity of the graph to reduce computational and storage costs, while enhancing the specificity of functional connections. The adjacency matrix A of thresholding Γ is

$$A_{c_1,c_2} = \begin{cases} 1, \text{ if } \Gamma_{c_1,c_2} \geq \tau \\ 0, \text{ else} \end{cases} \tag{5}$$

where τ is the threshold. The hyperparameter τ controls the sparsity of the graph structure. To adapt to the distribution of connection strengths in the functional connectivity matrices of different individuals' whole-brain networks, this study employs the percentile method to determine the dynamic threshold τ. Specifically, we define the proportion of retained connections as r, then τ is the connection strength at the $(1-r)\times 100$th percentile in the functional connectivity matrix Γ. An excessively large τ would result in the graph structure being overly sparse, leading to information loss, while an excessively small τ would result in the inclusion of too many false connections in the graph structure.

According to A, we can gain the whole-brain functional graph $G = \{V, E\}$. Where V and E are the node set and edge set respectively. In G, each node represents a Region of Interest (ROI), and edges indicate functional connections between ROIs. Considering that the Graph Attention Network (GAT) is capable of globally aggregating information across the entire graph and is adaptable to different types of graph structures, we choose GAT as the spatial feature extractor. The adaptive attention mechanism in GAT allows each node to dynamically allocate different weights to neighboring nodes during the information propagation process, which is advantageous for analyzing complex Brain Functional Connectivity (BFC). The embedding features of spatial-path are

$$Z_{\text{emb}}^S = \varepsilon_S(\mathcal{V}, E) \tag{6}$$

where ε_S is GAT network. $Z_{\text{emb}}^S \in \mathbb{R}^{n\times d_e}$, d_e is embedding dimensions. $\mathcal{V}_c = \Gamma_c$ is the embedding feature of V_c.

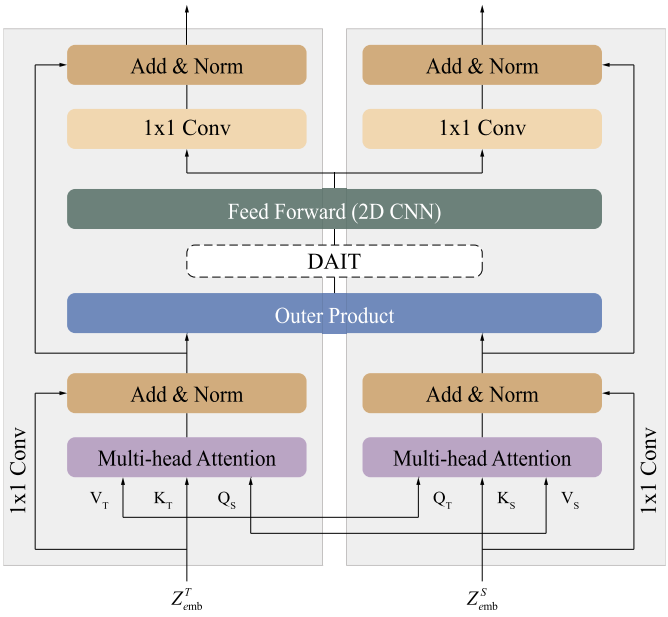
3.4. Multi-head dual-path cross-attention

We propose a spatiotemporal dual-path cross-attention block (DCA) to interact and integrate the complex spatiotemporal patterns in fMRI. The Structure of Dual-Path Cross-Attention Block (DCA) as shown in Fig. 3. Our DCA block also includes two steps: computing attention and feedforward. In the cross-attention stage, we interact spatial-temporal features by interleaving query and key vectors of dual-path. In the feedforward stage, we propose merging dual-path attention into spatiotemporal dual-path attention integrated tensor (DAIT) and then utilizing 2D convolution for nonlinear transformation to enhance the model's ability to co-represent spatiotemporal patterns.

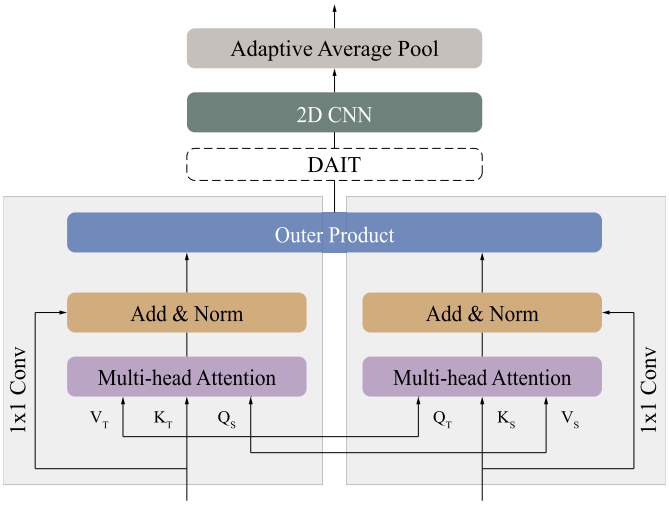
3.4.1. Cross-attention layer

Spatiotemporal dual-path cross-attention captures long-range spatiotemporal dependencies in time series data and helps enhance the model's generalization ability to adapt to multi-center fMRI data. Furthermore, the deep interaction of spatiotemporal information can more robustly extract potential spatiotemporal patterns in fMRI data. It is advantageous for alleviating the issue of disease pattern instability caused by the heterogeneity of ASD.

Initially, the embedded features of the spatiotemporal dual paths, Z_{emb}^T and Z_{emb}^S, are subjected to linear transformations independently to obtain two sets of query, key, and value, denoted as $\{Q_T, K_T, V_T\}$ and $\{Q_S, K_S, V_S\}$, respectively. Where, the dimensions of the query, key, and value for the temporal path are all $m \times d_e$, while for the spatial path are all $n \times d_e$.



(a) Non-last block



(b) Last block

Fig. 3. The structure of dual-path cross-attention block (DCA).

For temporal attention, we exchange the key vector K to intergrating the temporal and spatial pattens, as

$$\alpha_T\left(Q_S, K_T, V_T\right) = \text{softmax}\left(\frac{Q_S, (K_T)^T}{\sqrt{\delta_T}}\right) V_T, \tag{7}$$

$\alpha_T \in \mathbb{R}^{n \times d_e}$ Similarly, for spatial attention,

$$\alpha_S\left(Q_T, K_S, V_S\right) = \text{softmax}\left(\frac{Q_T, (K_S)^T}{\sqrt{\delta_S}}\right) V_S, \tag{8}$$

$\alpha_S \in \mathbb{R}^{m \times d_e}$.

Due to the different dimensions of Z_{emb} and α, the residual connections in STDCformer first undergo dimension alignment using 1×1 convolution before being added together.

3.4.2. Feed forward layer

To prevent the loss and distortion of spacetime information, we propose integrating dual-path features into a tensor while preserving the spacetime dimensions. Additionally, we employ two-dimensional convolution for forward processing to enhance spatiotemporal correlation analysis at a local scale. Initially, α_S and α_T are synthesized into the spatiotemporal dual-path attention integrated tensor (DAIT) through

$$\mathcal{A} = \alpha_S \times \alpha_T, \tag{9}$$

$\mathcal{A} \in \mathbb{R}^{m \times n \times d_e}$. $\times$ is outer product. $\mathcal{A}$ simultaneously preserves both the temporal and spatial dimensions. $\mathcal{A}$ comprehensively integrates spatiotemporal patterns and is utilized in downstream feedforward layers. To enhance the interaction of spatiotemporal local features, the feedforward layers in this work employ 2D convolution to accommodate the dimensions of spatiotemporal attention after concatenation.

$$\mathcal{Z} = \text{Conv}_{2\text{D}}(\mathcal{A}), \tag{10}$$

$\mathcal{Z} \in \mathbb{R}^{m \times n \times d_e}$. The d_e dimension is used as a channel in the STDCformer, similar to other feedforward layers. Like other feedforward layers, the input and output dimensions of the feedforward layer in STDCformer remain consistent, i.e., d_e.

For the stacked DCA block, the spatial and temporal dimensions of $\mathcal{Z}$ are respectively compressed by 1×1 convolutions and serve as the input for the temporal and spatial paths of the subsequent layer's DCA, as illustrated in Fig. 3 (a). For the final layer of DCA, DAIT adaptively pools the $1 \times d_e$ input for the prediction head, as depicted in Fig. 3 (b).

3.5. Prediction head

We obtain the predicted label $\hat{y}$ from $\mathcal{Z}$ through a multi-layer perceptron (MLP). The training process is supervised by the ground truth diagnostic labels y through minimizing the cross-entropy loss $\mathcal{L}$,

$$\mathcal{L} = \text{CrossEntropy}(\bar{y}, y) \tag{11}$$

3.6. Implementation

To mitigate the issues of vanishing and exploding gradients for more stable training of STDCformer, we employ the He initialization method, particularly suited for ReLU activation functions, to initialize the network weights, with biases initialized to zero. The Adaptive Moment Estimation (Adam) optimizer is utilized for optimizing the parameters of STDCformer, as its adaptive learning rate capability aids in faster convergence of the model. The configuration of tunable hyperparameters within the STDCformer architecture is grounded in experimentation.

4. Experiments and result analysis

4.1. Datasets

We evaluated STDCformer on Autism Brain Imaging Data Exchange (ABIDE) [25]. The ABIDE-I database consists of rs-fMRI and phenotypic data collected from 1112 subjects across 17 sites. In this study, we employed a subset of the data preprocessed using the C-PAC pipeline, which included 871 samples. We utilized 90 brain regions defined as ROIs using the Automated Anatomical Labeling (AAL) atlas. The time series lengths varied across different sites, ranging from 78 to 296. We excluded data from the site with the shortest time series length (78 temporal points). Additionally, we truncated the sequences of all remaining subjects based on the shortest remaining sequence length. We also excluded subjects with missing values. In summary, we utilized data from 836 subjects across 16 sites, with each subject having 116 temporal points of 90 ROIs. Among these 836 subjects, there were 386 ASD patients and 450 healthy controls.

4.2. Implementation details

Our STDCformer is implemented in PyTorch. All experiments were conducted on a PC equipped with an Intel Core i7-12700K CPU and an Nvidia RTX 3090 GPU.

In this work, the optimizer used is Adaptive Moment Estimation (Adam) with an initial learning rate of 1e-3. The learning rate is dynamically adjusted by ReduceLROnPlateau based on the validation loss, with a patience of 10 epochs and a minimum learning rate of 1e-6. The number of epochs is set to 100, with a batch size of 64. In STDCformer, the embedding dimensions for both the temporal and spatial pathways are set to 32. In the temporal path embedding layer, there are 2 layers of LSTM, and the parameter o_p of PPE is set to randomly vary within ± 0.005. In the spatial path embedding layer, there are 3 layers of GAT. In DCA, there are 8 attention heads for multi-head self-attention on each pathway, and the feedforward layer is ResNet18. The number of DCAs is 2. The classification head consists of a single-layer fully connected network.

Table 1
Comparison with baseline methods.

Method	ACC	AUC	SEN	SPE	F1
ResNet-1D	0.6265	0.6207	0.5427	0.6987	0.5717
Dlinear	0.6333	0.6088	0.4818	0.7358	0.4914
TCN	0.5595	0.5590	0.5513	0.5667	0.5375
LSTM	0.6131	0.6132	0.6154	0.6111	0.5963
GCN	0.6429	0.6214	0.3636	**0.8791**	0.4828
GAT	0.6369	0.6355	0.6154	0.6556	0.6115
Transformer [9]	0.6032	0.5830	0.6025	0.5635	0.5837
iTransformer [29]	0.5979	0.5875	0.4474	0.7275	0.4987
Crossformer [30]	0.6587	0.5892	0.3005	0.8778	0.3513
Autoformer [31]	0.6372	0.6128	0.5083	0.7174	0.5316
PatchTST [32]	0.6252	0.6019	0.6046	0.5992	0.5836
STDCformer	**0.6905**	**0.7017**	**0.7143**	0.6703	**0.6790**

We employ 5-fold stratified cross-validation for evaluation. The quantitative metrics utilized are accuracy (ACC), area under the curve (AUC), sensitivity (SEN), specificity (SPE), and F1 score (F1).

4.3. Comparison experiences

4.3.1. Comparison with baseline methods

STDCformer is compared with a series of baseline methods. STDCformer is compared against a series of baseline methods, including: ResNet-1D and Dlinear [26]; TCN [27] and LSTM [3] widely used for time series analysis; GCN [28] and GAT [10] for the graph convolution paradigm; a series of X-formers. The results are reported in Table 1.

The Table 1 demonstrates that our STDCformer exhibits the best performance in terms of ACC, AUC, SEN, and F1. In specificity, models such as GCN, GAT, iTransformer, Autoformer, Pyraformer, and Crossformer outperform STDCformer, but they also exhibit significantly low sensitivity.

Within the group of TCN vs. LSTM, LSTM consistently outperforms TCN. This phenomenon may be due to the gate mechanism of LSTM, which flexibly learns and maintains the time series internal structure and patterns. In contrast, TCN is better at capturing short-term dependencies due to the fixed convolution kernel size limitation. Therefore, LSTM is better suited for data like fMRI, which involves multicenter biases and individualized differences. Thus, we chose LSTM as the token embedding model for the temporal path.

Within the group of GCN vs. GAT, GAT significantly outperforms GCN in balancing sensitivity and specificity. This phenomenon may be due to the node-level attention mechanism in GAT, which can more flexibly capture relationships between nodes. Furthermore, the sparsity in calculating attention scores in GAT helps the model focus more on important information. This characteristic enhances the robustness of the model to multi-center biases and individual differences in fMRI data. Thus, we chose GAT as the feature embedding model for the spatial path.

Within the X-former group, it is interesting to concentrate that models such as iTransformer and Crossformer, which emphasize the interaction of dimensional information at the spatial scale, perform less well than the vanilla Transformer. This phenomenon suggests that capturing the correlation of neural activity between brain regions from fMRI data under the self-attention paradigm is challenging. On the contrary, models like Autoformer and PatchTST, which further emphasize temporal scale features, show performance improvements. In summation, improvements focusing on the dynamic nature of fMRI signals and efforts to introduce more powerful spatial information extraction models (such as graph convolution) are allworth considering.

4.3.2. Comparison with the state-of-the-art methods

STDCformer is compared with a series of state-of-the-art (SOTA) methods. The competing methods include:

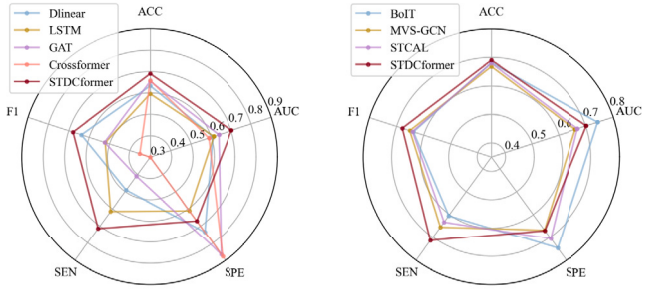
- Method focusing on temporal patterns in the Transformer paradigm: BolT [5].
- Methods emphasizing spatial patterns in the graph convolution paradigm: BrainGNN [7], MVS-GCN [33].
- Methods that simultaneously highlight spatiotemporal patterns: TCN-GNN [11], ST-Transformer [12], ST-GCN [13], BrainTGL [15], STA-GIN [34], STCAL [14].

The results are reported in Table 2. Unmarked competing methods indicate implementation based on publicly available code in the paper, and $*$ Indicates reproduction of the pipeline proposed in the paper.

In the methods that consider the spatiotemporal patterns in Table 2, only BrainTGL and STCAL exhibit performance comparable to Bolt, which emphasizes only temporal patterns, and MVS-GCN, which focuses solely on spatial patterns. In contrast to other methods considering spatiotemporal patterns, the common characteristic of BrainTGL and STCAL is the representation of both time and space patterns in different branches. This feature implies the rationality of adopting a dual-path architecture. Furthermore, the superior performance of Bolt compared to the vanilla Transformer underscores once again the necessity of specialized transformations tailored for the dynamics of time series. Additionally, within the spatial relationship group based on graph modeling, the performance

Table 2
Comparison with SOTA methods.

Method	ACC	AUC	SEN	SPE	F1
BolT	0.6814 (0.0360)	**0.7430** (0.0322)	0.6049 (0.1137)	0.7477 (0.1192)	0.6381 (0.0685)
BrainGNN	0.6420 (0.0426)	0.6159 (0.0399)	0.4641 (0.1607)	**0.7677** (0.1550)	0.5212 (0.0774)
MVS-GCN	0.6687 (0.0284)	0.6613 (0.0283)	0.6547 (0.1191)	0.6679 (0.0986)	0.6514 (0.0711)
TCN-GNN	0.6071 (0.0260)	0.5682 (0.0251)	0.4194 (0.1070)	0.7170 (0.1128)	0.4407 (0.0469)
STAGIN	0.6302 (0.0402)	0.6269 (0.0348)	0.5832 (0.1203)	0.6707 (0.1212)	0.5906 (0.0429)
ST-Transformer	0.6026 (0.0356)	0.5868 (0.0309)	0.4963 (0.0783)	0.6774 (0.1014)	0.5251 (0.0536)
ST-GCN	0.6111 (0.0400)	0.5921 (0.0436)	0.5153 (0.0937)	0.6689 (0.0694)	0.5362 (0.0720)
BrainTGL	0.6731 (0.0320)	0.6592 (0.0335)	0.6114 (0.0574)	0.7071 (0.1244)	0.6213 (0.0240)
STCAL*	0.6804 (0.0185)	0.6694 (0.0179)	0.6327 (0.0485)	0.7060 (0.0616)	0.6413 (0.0451)
STDCformer	**0.6905** (0.0124)	0.7017 (0.0101)	**0.7143** (0.0355)	0.6703 (0.0535)	**0.6790** (0.0337)



(a) Compare with baseline (b) Compare with SOTA

Fig. 4. Performance comparison of STDCformer and competitor.

improvement of MVS-GCN using multi-view graphs compared to BrainGNN suggests the crucial role of graph topology in identifying ASD.

Our method demonstrates an overall advantage in the standard deviation of performance across 5-fold data. This not only underscores the stability of our model but also partially reflects its superiority in adapting to data variances. Such superiority stems not only from the ability of our model's positional encoding to accommodate individual noise but also from its capability to learn more comprehensive and intrinsic spatiotemporal features through modules like cross-attention and DAIT, thus enabling it to adapt to local noise associated with individual differences.

4.3.3. Comprehensive analysis of comparison experiences

We compared the highest ACC models from each group of baseline and SOTA methods with our STDCformer in Fig. 4.

Fig. 4 illustrates the advantages of our STDCformer in terms of performance balance, especially in balancing sensitivity and specificity. This advantage is particularly prominent when compared to baseline methods. Additionally, our STDCformer demonstrates a significant advantage in sensitivity. Sensitivity measures the ability of the model to identify true positive cases. High sensitivity indicates that the model can better capture cases that truly have the disease. Therefore, high sensitivity contributes to the early detection and treatment of ASD. Furthermore, we observe that the performance of most SOTA methods surpasses that of baseline methods. This phenomenon indicates the significance and necessity of designing and adapting models specifically for fMRI data.

4.4. Ablation studies

4.4.1. Effectiveness of dual-path architecture

We demonstrate the effectiveness of the dual-path architecture through three sets of ablation experiments.

- The first set of experiments consists of ablating at the entire path level. We individually employ temporal path (T-Path) and spatial path (S-Path) to perform tasks. Since single paths cannot achieve cross-attention, the self-attention layer in this set of experiments is the same as that in a vanilla Transformer.
- The second set of experiments primarily demonstrated the effectiveness of the dual-path architecture of the embedding layer. We degraded the embedding layer to output a single embedding feature. This set comprised four experiments: retaining only the temporal embedding feature (T Embedding); retaining only the spatial embedding feature (S Embedding); concatenating the temporal embedding module and spatial embedding module sequentially (T-S Embedding); concatenating the spatial embedding

Table 3
Effectiveness of spatiotemporal dual-path framework.

Method	ACC	AUC	SEN	SPE	F1
T-Path	0.6429	0.6444	0.6667	0.6222	0.6341
S-Path	0.6587	0.6421	0.4286	0.8556	0.5366
T Embedding	0.6310	0.6342	0.6795	0.5889	0.6310
S Embedding	0.5893	0.5825	0.4872	0.6778	0.5241
T-S Embedding	0.6369	0.6355	0.6154	0.6556	0.6115
S-T Embedding	0.5928	0.5828	0.4545	0.7111	0.5072
T Attention	0.6683	0.6675	0.6588	0.6763	0.6458
S Attention	0.6071	0.6094	0.6364	0.5824	0.5976
STDCformer	0.6905	0.7017	0.7143	0.6703	0.6790

module and temporal embedding module sequentially (S-T Embedding). In T-S Embedding, the node features of graph convolution are the output of the temporal embedding module, and the temporal embedding module adopts a channel-wise independent approach. In S-T Embedding, the node features of graph convolution correspond to the time series. To maintain the dual-path architecture of the self-attention layer, we obtained two sets of Q, K, and V vectors through two sets of linear layers for cross-attention.

- The third set of experiments mainly demonstrates the effectiveness of cross-attention in the dual-path architecture. We retain cross-attention but degrade the model to have only single-path cross-attention. This set includes two experiments: introducing spatial features only in the query vectors in the temporal path (T Attention); introducing temporal features only in the query vectors in the spatial path (S Attention).

The results are reported in Table 3.

In Table 3, the superior performance of our STDCformer compared to all ablation experiments highlights the effectiveness of the dual-path architecture. Overall, experiments solely emphasizing spatial information perform less effectively compared to experiments solely emphasizing temporal information. This phenomenon is attributed to spatial pathways only emphasizing global dependency among variables while disregarding local dynamic details in temporal. Additionally, graph convolutions may suffer from over-smoothing, diminishing the modeling of dependencies among distant nodes. Hence, we propose supplementing spatial information within models that primarily model sequential information.

The T-Path and S-Path experiments exhibit advantages in sensitivity and specificity, respectively. This phenomenon underscores the necessity of integrating temporal and spatial information. It is noteworthy that the performance of experiments T-Path and S-Path, compared to state-of-the-art (SOTA) methods (Table 2), is not significantly inferior, and the comprehensive performance of experiment T-Path surpasses that of baseline methods (Table 1). This observation highlights the effectiveness of our modifications to the temporal path and spatial path.

Results from the second set of experiments, modifying the embedding layer architecture, are intriguing. Firstly, the performance of T Embedding and S Embedding, which embed single-path information but extend to dual-path attention, is inferior to that of experimental T-Path and S-Path. This result suggests that the information embedded in features may not be sufficient to support the learning of dual-path attention. Results from the second set of experiments, modifying the embedding layer architecture, are intriguing. Firstly, the performance of T Embedding and S Embedding, which embed single-path information but extend to dual-path attention, is inferior to that of experimental T-Path and S-Path. This is partly attributed to the exchange of information between single paths leading to partial loss or confusion of information during transmission, and partly to the overfitting risk and increased training complexity brought by modeling single-path embedding through cross-path attention. Secondly, both the enhancement of T-S Embedding over T Embedding and the enhancement of S-T Embedding over S Embedding are narrow. This phenomenon indicates the challenging nature of hybrid analysis of spatiotemporal patterns. Moreover, the significantly lower performance of S-T Embedding compared to T-S Embedding suggests the potential distortion of information in an entangled analysis of spatiotemporal patterns. Simultaneously, this result indicates that the structure of connecting temporal and spatial feature extraction modules within a single path is inflexible. Cross groups, the performance degradation caused by experimental T-S Embedding and S-T Embedding is more severe than the performance degradation caused by ablating self-attention layers. This phenomenon confirms the importance of embedding temporal and spatial information in different paths.

To further elucidate the efficacy of the dual-path architecture we designed, we conducted Leave-One-Site-Out (LOSO) cross-validation on the 16 sites adopted in this study. The accuracy across the 16 sites is illustrated in Fig. 5. From Fig. 5, it can be observed that our dual-path architecture exhibits advantages in generalization across most sites. It is noteworthy that the generalization of the temporal pathway (T-Path) is overall inferior to that of the spatial pathway (S-Path). This phenomenon arises because spatial patterns consider the relative relationships between signals among brain regions rather than the dynamics of the signals themselves, thus being less sensitive to noise. This underscores the importance of considering the whole-brain functional connectivity patterns. It is noteworthy that, in comparison to spatial paths, although our method introduces temporal paths, it is not further affected by the sensitivity of temporal patterns to differences between sites, but rather enhances generalization ability. This phenomenon indicates that our model captures effective features for ASD identification from signals.

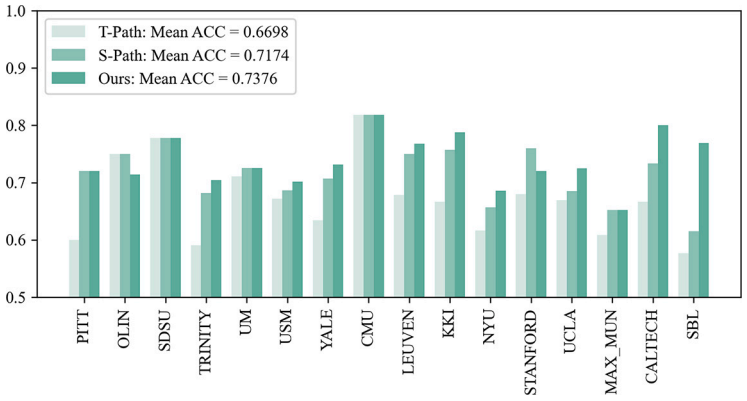


Fig. 5. The effectiveness of dual-path architecture in enhancing model generalization.

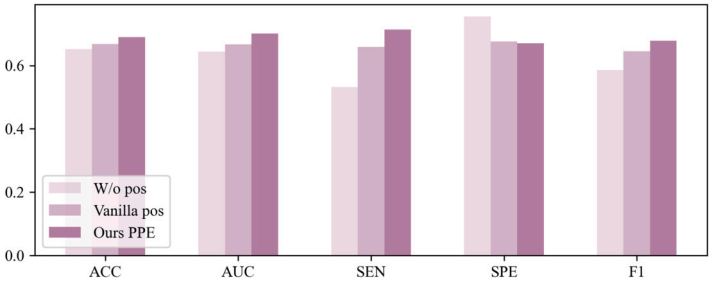


Fig. 6. The impact of PPE.

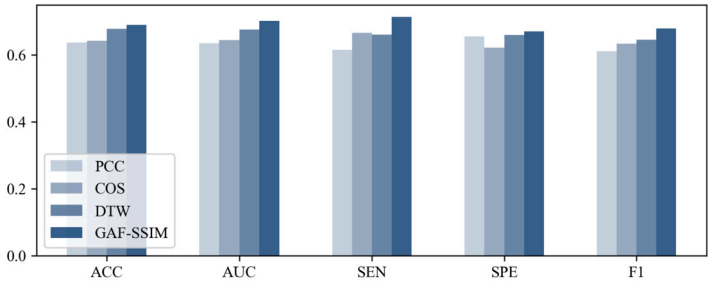


Fig. 7. The impact of GAF-SSIM.

4.4.2. Ablation studies on temporal embedding layer

In the temporal embedding layer, the most significant modification in this work is the introduction of Perturbed Position Encoding (PPE). We demonstrate the effectiveness of PPE by comparing it with no position encoding (W/o pos) and the commonly used sinusoidal position encoding (Vanilla pos). The results are in Fig. 6.

The results in Fig. 6 indicate that our PPE contributes to enhancing the model's ability to recognize ASD. Specifically, PPE enhances the sensitivity of the model. This is attributed to the introduction of stochastic noise in PPE, which increases the randomness of the model, thereby emphasizing the extraction of primary features from the data and improving the model's generalization capability. Furthermore, the improvements in AUC and F1 brought about by PPE are noteworthy. By adding stochastic noise to the positional encoding, the model exhibits greater robustness to minor variations in the input data, thus better accommodating individual differences. Additionally, stochastic noise can be seen as a form of data augmentation and regularization, aiding in reducing the risk of model overfitting.

4.4.3. Ablation studies on spatial embedding layer

In the spatial embedding layer, the primary contribution of this work is proposing to establish whole-brain functional connectivity (BFC) based on GAF-SSIM. We demonstrate the effectiveness of constructing BFC using GAF-SSIM by comparing it with Pearson correlation coefficient (PCC), cosine similarity (COS), and dynamic time warping (DTW). The results are in Fig. 7.

Fig. 7 demonstrates that the BFC constructed by our GAF-SSIM exhibits the best performance on all metrics. In particular, the BFC built by GAF-SSIM shows significant superiority in terms of SEN. Compared to the BFC constructed via PCC, the BFC constructed via COS disregards the influence of signal amplitude, thus showing better performance in cosine similarity when dealing with significant

Table 4
Effectiveness of DCA.

Method		ACC	AUC	SEN	SPE	F1
W/o Cross		0.6826	0.6905	0.7922	0.5889	0.6971
W/o DAIT	Fusion	0.6131	0.6132	0.6154	0.6111	0.5963
	T-Dim	0.6369	0.6355	0.6154	0.6556	0.6115
	S-Dim	0.6108	0.5995	0.4545	0.7444	0.5185
STDCformer		0.6905	0.7017	0.7143	0.6703	0.6790

individual differences in BOLD signals. The BFC constructed via DTW shows markedly enhanced performance in terms of ACC and AUC. DTW robustly measures the lag and shape differences between two time series by elastically aligning local segments of the time series, enhancing its robustness against global noise [35]. This implies that DTW can adapt to deviations between different sites to a certain extent. However, DTW is sensitive to local noise of individuals, and local alignment may lead to the neglect or amplification of minor signal changes, thus affecting the identification of key features. Therefore, the BFC constructed via DTW exhibits lower sensitivity compared to the BFC constructed via COS. In contrast, our proposed GAF-SSIM combines the phase and amplitude information of time series to capture the global nonlinear relationships of temporal information, enhancing cross-individual consistency. Furthermore, GAF-SSIM provides a higher resolution feature representation that assists in accurately reflecting the complex relationships between brain regions.

4.4.4. Ablation studies on dual-path cross-attention

We demonstrate the effectiveness of self-attention and feedforward in the dual-path cross-attention (DCA) block through two groups of ablation experiments. In the first group (W/o Cross), we omitted cross-attention and instead calculated the attention for temporal and spatial paths separately in parallel, then integrated the attention from both branches into a Dual-Path Integrated Tensor (DAIT) for feedforward processing. The second group (W/o DAIT) comprises three experiments: (Fusion) DAIT is not used, where temporal path attention vectors and spatial path attention vectors are encoded into the same dimension by two linear layers, then the dual-path features are fused into one feature using softmax gating and feedforward through 1D convolution before being input to the prediction head; (T-Dim) Only the temporal dimension of DAIT is retained (with spatial dimension averaged), and fed through one-dimensional convolution before being input to the prediction head; (S-Dim) Only the spatial dimension of DAIT is retained (with temporal dimension averaged), and fed through one-dimensional convolution before being input to the prediction head. Note that, to ensure the stacking of DCA blocks, in the second set of three experiments, we only modified the last DCA. The results are in Table 4.

Table 4 demonstrates the effectiveness of DCA in our STDCformer. Additionally, it reveals that DAIT significantly impacts model performance more than cross-attention. When retaining information from both paths and only omitting interleaved attention (W/o Cross), the accuracy decreases by only 0.79%. However, despite the high SEN of W/o Cross, its SPE is low. Considering the trade-off between SEN and SPE, the advantage of cross-attention should still be emphasized. The performance of the three experiments in the W/o DAIT group significantly deteriorates compared to the STDCformer with DAIT. Special note that Fusion performs worse than using only the time dimension T-Dim. This phenomenon offers that even in deep layers, there exists considerable heterogeneity in information from both paths. Furthermore, it implies that temporal information complements spatial information. Therefore, both the dimension of time and the dimension of space should be preserved simultaneously.

Furthermore, we present in Fig. 8 the AUC-ROC curves of the experiments in Table 4 to visually demonstrate that our DCA outperforms the configurations of other ablation experiments in terms of overall performance.

5. Discussion

5.1. Limitations of STDCformer

STDCformer still has some gaps that can be improved. In terms of model architecture, our approach relies on a single template to partition brain regions, which may introduce bias when applied universally across all subjects. Therefore, in future work, we intend to explore extracting features from multiple templates and further enhancing model stability through consistency comparisons between templates. Regarding data input, STDCformer embeds temporal and spatial features through dual pathways. Thus, we will consider decomposing signals into components such as trend and seasonality for targeted learning of spatiotemporal patterns, which could potentially aid the model in capturing spatiotemporal patterns more effectively.

Furthermore, our current method lacks learnable parameters for integrating spatiotemporal features into DAIT, which may limit the model's ability to capture deeper dependencies between spatiotemporal features. Future efforts will involve considering adaptive fusion techniques, such as gating mechanisms, to potentially enhance the effectiveness of features for ASD identification. Additionally, this study did not incorporate non-imaging information such as demographic or clinical data. Integrating non-imaging information not only introduces more information beneficial for precise ASD identification but also assists the model in capturing potential associations between individual characteristics and imaging phenotypes, thus facilitating interpretative analyses of the model. Our next steps will explore strategies for appropriately integrating non-imaging information into deep learning models.

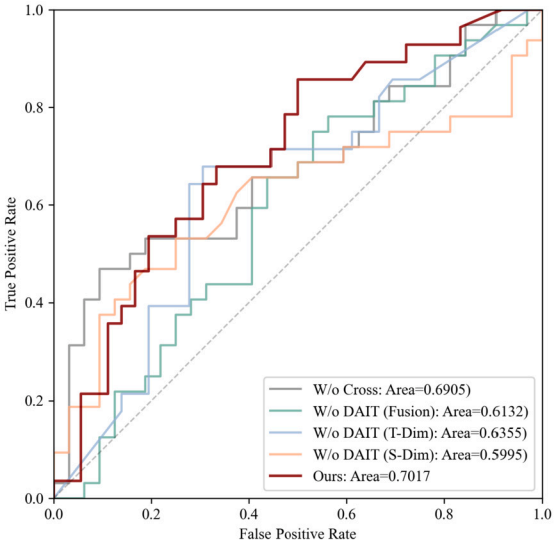


Fig. 8. Verify the effectiveness of DCA through the AUC-ROC curve.

Table 5
Top-10 importance ROIs for ASD identification.

Rank	Map	Abbreviation	Description
1		Frontal_Sup_Orb_R	Right Superior Frontal Gyrus, Orbital
2		Heschl_L	Left Heschl's gyrus
3		Postcentral_L	Left Parietal Lobe
4		Cingulum_Ant_R	Right Anterior Cingulate Gyrus
5		Fusiform_R	Right Fusiform Gyrus
6		Temporal_Inf_R	Right Inferior Temporal Gyrus
7		Frontal_Sup_L	Left Superior Frontal Gyrus
8		Temporal_Pole_Sup_R	Right Superior Temporal Pole
9		Temporal_Inf_L	Left Inferior Temporal Gyrus
10		Calcarine_R	Right Calcarine Sulcus

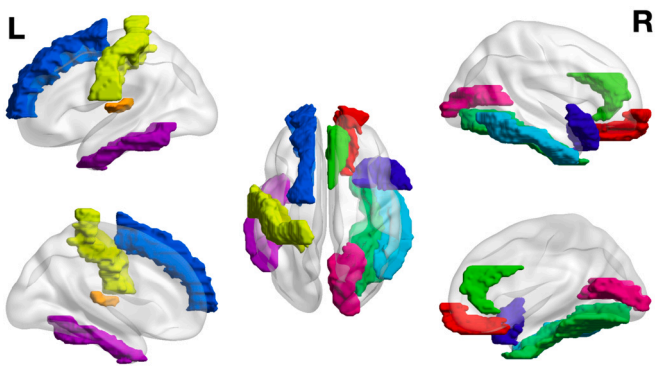


Fig. 9. Visualization of top-10 importance ROIs for ASD identification.

5.2. Rs-fMRI-derived neuroimaging markers

We extracted the ROI importance ranking from STDCformer based on Shapley Additive exPlanations (SHAP) [36]. The fMRI-derived top-10 important brain regions are reported in Table 5 and visualized in Fig. 9.

The top-10 important brain regions (as shown in Table 5 and Fig. 9) revealed by STDCformer associated with ASD identification are largely consistent with those reported in the existing literature, as shown in Table 6. Please note that the clues we provide are very preliminary and not rigorous. However, the preliminary clues provided in this article may be somewhat helpful for ASD research. In the future, we will further validate these clues on a larger scale and more comprehensive data.

Table 6
References to our recommended ROIs.

Area	ROI	References
Frontal lobe	Frontal_Sup_Orb_R Frontal_Sup_L	[37][38][39]
Parietal lobe	Postcentral	[40]
Temporal lobe	Heschl_L Fusiform_R Temporal_Inf_R Temporal_Pole_Sup_R Temporal_Inf_L	[41][42][43][44][45]
Occipital lobe	Calacarine_R	[46]
Cingulate gyrus	Cingulum_Ant_R	[47][48][49]

5.2.1. Anomalies in regulating negative emotions may be one of the symptoms in individuals with ASD

In Table 5, the regions primarily involved in emotion regulation, such as Frontal_Sup_Orb and Cingulum_Ant, both show activity on the right side. The right-sided regions of the frontal lobe and cingulate gyrus are predominantly associated with the regulation of negative emotions. Meanwhile, the right Temporal_Inf and Fusiform, which are involved in emotion regulation and facial expression recognition, also showed abnormalities. Abnormal neural activity in Frontal_Sup_Orb_R and Cingulum_Ant_R may suggest potential clues regarding behaviors such as anxiety, anger, and feelings of loneliness in individuals with ASD.

5.2.2. Abnormalities in language may be one of the symptoms in individuals with ASD

In Table 5, both the left Heschl, left Temporal_Inf and left Frontal_Sup regions demonstrate relatively high levels of contribution. For more typical right-handed individuals, the left Heschl and left Temporal_Inf regions are proximal to Wernicke's area in the temporal lobe, and left Frontal_Sup region is proximal to Broca's area in the frontal lobe. Wernicke's area and Broca's area together form the language hub. Anomalies in Heschl_L and Frontal_Sup_L may suggest potential clues regarding language-related impairments in individuals with ASD.

5.2.3. Anomalies in sensation may be one of the symptoms in individuals with ASD

In Table 5, the auditory cortex Heschl, somatosensory cortex Postcentral, and visual cortex Calcarine regions exhibit higher levels of contribution. Furthermore, we examined the feature effects in SHAP. The results indicate that neural activation levels in the Heschl and Calcarine regions are lower in ASD patients, while the activation level in the Postcentral region is higher. The abnormalities in the Heschl and Calcarine regions may suggest difficulties in auditory and visual information processing in ASD patients. The heightened activation level in the Postcentral region, located in the parietal lobe of the brain, may imply increased sensitivity to tactile stimuli in ASD patients. Their somatosensory sensitivity may be related to a preference for activities involving pressure touch, such as hugging or gentle tapping. In addition, these sensory abnormalities in individuals with autism spectrum disorder may be one of the potential clues as to why they struggle to form new behavioral patterns and resort to repetitive stereotypical behaviors.

5.3. Implication of findings

The STDCformer extension we propose expands the current applications of deep learning in neural imaging analysis and demonstrates the potential of deep learning techniques in handling complex medical image data.

By identifying brain regions associated with ASD from the STDCformer, the transparency of the model's decision-making process is enhanced, providing a more diverse perspective for understanding the neural basis of ASD and exploring its underlying biological mechanisms. While there remains a considerable gap between these findings and real-world insights into ASD, this demonstrates the potential value of AI-guided neuroimaging biomarker development for ASD diagnosis and the discovery of new therapeutic targets.

Our approach maintains good stability while achieving competitive performance, opening new possibilities for personalized and precise ASD diagnosis using deep learning. Furthermore, the robustness of our STDCformer in multicenter data suggests its potential for broader and more diverse applications in various clinical settings.

From the perspective of model architecture, the STDCformer offers a novel deep learning framework designed to handle both temporal and spatial data. This framework may be applicable to other fields that require simultaneous consideration of temporal dynamics and spatial relationships.

6. Conclusion

This work proposes a spatial-temporal dual-path cross-attention model named STDCformer for identifying ASD based on fMRI data. STDCformer deeply interacts and integrates spatial-temporal pattern information while retaining domain-specific information in both time and space. We anticipate extending our work to additional domains. The dual-path architecture, perturbation position encoding, GAF image similarity-based correlation measurement, and dual-path attention integrated tensor modules proposed in STDCformer

have generality. Thus, our work may potentially assist the model of the multi-head self-attention paradigm in various scenarios. Moreover, neuroimaging biomarkers derived from STDCformer through interpretable techniques could offer clues to understanding the physiological mechanisms of ASD. In conclusion, our work once again demonstrates the potential of deep learning technology in ASD research.

CRediT authorship contribution statement

Haifeng Zhang: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Chonghui Song:** Writing – review & editing, Supervision, Conceptualization. **Xiaolong Zhao:** Visualization, Software. **Fei Wang:** Visualization, Validation. **Yunlong Qiu:** Software. **Hao Li:** Software. **Hongyi Guo:** Software.

Declaration of competing interest

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

Data availability

The data associated with this study has not been deposited into a publicly available repository. However, the data contained in this article are available for inspection and use according to the referenced public dataset links. We encourage other researchers to evaluate our findings and build upon our work using the provided data.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61773006 and Medical-Industrial Intersection Joint Foundation of Liaoning Province under Grant 2022-YGJC-14.

References

[1] H.N. Ingabire, et al., Stability analysis of fMRI BOLD signals for disease diagnosis, IEEE Trans. Neural Syst. Rehabil. Eng. 30 (Apr. 2022) 967–978.
[2] P. Moridian, et al., Automatic autism spectrum disorder detection using artificial intelligence methods with MRI neuroimaging: a review, Front. Mol. Neurosci. 15 (Oct. 2022).
[3] N.C. Dvornek, P. Ventola, K.A. Pelphrey, J.S. Duncan, Identifying autism from resting-state fMRI using long short-term memory networks, in: Machine Learning in Medical Imaging (MLMI), 2017.
[4] W. Jiang, et al., CNNG: a convolutional neural networks with gated recurrent units for autism spectrum disorder classification, Front. Aging Neurosci. 14 (Jul. 2022).
[5] H.A. Bedel, I. Sivgin, O. Dalmaz, S.U.H. Dar, T. Çukur, BolT: fused window transformers for fMRI time series analysis, Med. Image Anal. 88 (Aug. 2023).
[6] S. Parisot, et al., Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease, Med. Image Anal. 48 (Aug. 2018) 117–130.
[7] X. Li, et al., BrainGNN: interpretable brain graph neural network for fMRI analysis, Med. Image Anal. 74 (Dec. 2021).
[8] R. Yu, C. Pan, X. Fei, M. Chen, D. Shen, Multi-graph attention networks with bilinear convolution for diagnosis of schizophrenia, IEEE J. Biomed. Health Inform. 27 (3) (Mar. 2023) 1443–1454.
[9] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: Proc. Int. Conf. Learn. Represent. (ICLR), 2018, pp. 1–12.
[11] T. Azevedo, et al., A deep graph neural network architecture for modelling spatio-temporal dynamics in resting-state functional MRI data, Med. Image Anal. 79 (July 2022).
[12] X. Deng, J. Zhang, R. Liu, K. Liu, Classifying ASD based on time-series fMRI using spatial–temporal transformer, Comput. Biol. Med. 151, Part B (Dec. 2022).
[13] S. Gadgil, Q. Zhao, A. Pfefferbaum, E.V. Sullivan, E. Adeli, K.M. Pohl, Spatio-temporal graph convolution for resting-state fMRI analysis, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI), vol. 12267, 2020, pp. 528–538.
[14] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, K.C. Tan, Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state fMRI data, IEEE Trans. Neural Netw. Learn. Syst. (Feb. 2023), early access.
[15] L. Liu, et al., BrainTGL: a dynamic graph representation learning model for brain network analysis, Comput. Biol. Med. 153 (Feb. 2023).
[16] K.-W. Park, S.-B. Cho, A residual graph convolutional network with spatio-temporal features for autism classification from fMRI brain images, Appl. Soft Comput. 142 (July 2023).
[17] J. Zhang, L. Zhou, L. Wang, M. Liu, D. Shen, Diffusion kernel attention network for brain disorder classification, IEEE Trans. Med. Imaging 41 (10) (Oct. 2022) 2814–2827.
[18] X. Xing, et al., Dynamic spectral graph convolution networks with assistant task training for early MCI diagnosis, in: MICCAI 2019: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, vol. 11767, Oct. 2019, pp. 639–646.
[19] W. Cui, et al., Personalized functional connectivity based spatio-temporal aggregated attention network for MCI identification, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (Apr. 2023) 2257–2267.
[20] J. Liu, et al., Effective hyper-connectivity network construction and learning: application to major depressive disorder identification, Comput. Biol. Med. 171 (Mar. 2024).
[21] Y. Li, J. Liu, Z. Tang, B. Lei, Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for MCI identification, IEEE Trans. Med. Imaging 39 (9) (Sept. 2020) 2818–2830.
[22] Y. Li, J. Liu, Y. Jiang, Y. Liu, B. Lei, Virtual adversarial training-based deep feature aggregation network from dynamic effective connectivity for MCI identification, IEEE Trans. Med. Imaging 41 (1) (Jan. 2022) 237–251.
[23] J. Liu, et al., Deep fusion of multi-template using spatio-temporal weighted multi-hypergraph convolutional networks for brain disease analysis, IEEE Trans. Med. Imaging 43 (2) (Feb. 2024) 860–873.

[24] Z. Wang, T. Oates, Imaging time-series to improve classification and imputation, in: Proc. 24th Int. Joint Conf. Artif. Intell (IJCAI), July 2015, pp. 3939–3945.
[25] A.D. Martino, et al., The autism brain imaging data exchange: towards large-scale evaluation of the intrinsic brain architecture in autism, Mol. Psychiatry 19 (6) (Jun. 2014).
[26] A. Zeng, M. Chen, L. Zhang, Q. Xu, Are transformers effective for time series forecasting?, in: Proc. AAAI Conf. Artif. Intell., 2023.
[27] C. Lea, R. Vidal, A. Reiter, G.D. Hager, Temporal convolutional networks: a unified approach to action segmentation, in: Proc. Eur. Conf. Comput. Vis. (ECCV), vol. 9915, 2016, pp. 47–54.
[28] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proc. Int. Conf. Learn. Represent. (ICLR), 2017.
[29] Y. Liu, et al., iTransformer: Inverted Transformers Are Effective for Time Series Forecasting, arXiv, 2023.
[30] Y. Zhang, J. Yan, Crossformer: transformer utilizing cross-dimension dependency for multivariate time series forecasting, in: Proc. Int. Conf. Learn. Represent. (ICLR), 2023.
[31] H. Wu, J. Xu, J. Wang, M. Long, Autoformer: decomposition transformers with auto-correlation for long-term series forecasting, Adv. Neural Inf. Process. Syst. 34 (2021) 22419–22430.
[32] Y. Nie, N.H. Nguyen, P. Sinthong, J. Kalagnanam, A time series is worth 64 words: long-term forecasting with transformers, in: Proc. Int. Conf. Learn. Represent. (ICLR), 2023.
[33] G. Wen, P. Cao, H. Bao, W. Yang, T. Zheng, O. Zaiane, MVS-GCN: a prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis, Comput. Biol. Med. 142 (Mar. 2022).
[34] B.-H. Kim, J.C. Ye, J.-J. Kim, Learning dynamic graph representation of brain connectome with spatio-temporal attention, Adv. Neural Inf. Process. Syst. 34 (2021) 4314–4327.
[35] A.C. Linke, et al., Dynamic time warping outperforms Pearson correlation in detecting atypical functional connectivity in autism spectrum disorders, NeuroImage 223 (Dec. 2020).
[36] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017) 4765–4774.
[37] K.A.R. D. -Thomas, et al., Effects of age and symptomatology on cortical thickness in autism spectrum disorders, Res. Autism Spectr. Disord. 7 (1) (Jan. 2013).
[38] S. Xu, et al., Altered functional connectivity in children with low-function autism spectrum disorders, Front. Neurosci. 13 (Aug. 2019).
[39] X.-W. Zhu, L.-L. Zhang, Z.-M. Zhu, L.-Y. Wang, Z.-X. Ding, X.-M. Fang, Altered intrinsic brain activity and connectivity in unaffected parents of individuals with autism spectrum disorder: a resting-state fMRI study, Front. Human Neurosci. 16 (30) (Sep. 2021).
[40] E.T. Klapwijk, et al., Differential fairness decisions and brain responses after expressed emotions of others in boys with autism spectrum disorders, J. Autism Dev. Disord. 47 (May 2017) 2314–2329.
[41] D. Kim, et al., Overconnectivity of the right Heschl's and inferior temporal gyrus correlates with symptom severity in preschoolers with autism spectrum disorder, Autism Res. 14 (11) (Sep. 2021) 2314–2329.
[42] D. Hubl, et al., Functional imbalance of visual pathways indicates alternative face processing strategies in autism, Neurology 61 (9) (Nov. 2003) 1232–1237.
[43] A.D. Martino, K. Ross, L.Q. Uddin, A.B. Sklar, F.X. Castellanos, M.P. Milham, Functional brain correlates of social and nonsocial processes in autism spectrum disorders: an activation likelihood estimation meta-analysis, Biol. Psychiatry 65 (1) (Jan. 2009) 63–74.
[44] X. Yue, et al., Regional dynamic neuroimaging changes of adults with autism spectrum disorder, Neuroscience 523 (Jul. 2023) 132–139.
[45] S.-Y. Kim, et al., Abnormal activation of the social brain network in children with autism spectrum disorder: an fMRI study, Psychiatr. Investig. 12 (1) (Oct. 2014) 37–45.
[46] P.S. Lee, et al., Atypical neural substrates of embedded figures task performance in children with autism spectrum disorder, NeuroImage 38 (1) (Oct. 2007) 184–193.
[47] J. Hau, et al., The cingulum and cingulate U-fibers in children and adolescents with autism spectrum disorders, Hum. Brain Mapp. 40 (11) (Aug. 2019) 3153–3164.
[48] S.M. Haigh, T.A. Keller, N.J. Minshew, S.M. Eack, Reduced white matter integrity and deficits in neuropsychological functioning in adults with autism spectrum disorder, Autism Res. 13 (5) (May. 2020).
[49] Z.K. K. -Reza, M.A. Shahram, H. Zare, Altered resting-state functional connectivity of the brain in children with autism spectrum disorder, Radiol. Phys. Technol. 16 (Apr. 2023) 284–291.