Spatial–Temporal Co-Attention Learning for Diagnosis of Mental Disorders From Resting-State fMRI Data

Rui Liu, Zhi-An Huang, Yao Hu, Zexuan Zhu, *Senior Member, IEEE*, Ka-Chun Wong, and Kay Chen Tan, *Fellow, IEEE*

***Abstract*— Neuroimaging techniques have been widely adopted to detect the neurological brain structures and functions of the nervous system. As an effective noninvasive neuroimaging technique, functional magnetic resonance imaging (fMRI) has been extensively used in computer-aided diagnosis (CAD) of mental disorders, e.g., autism spectrum disorder (ASD) and attention deficit/hyperactivity disorder (ADHD). In this study, we propose a spatial–temporal co-attention learning (STCAL) model for diagnosing ASD and ADHD from fMRI data. In particular, a guided co-attention (GCA) module is developed to model the intermodal interactions of spatial and temporal signal patterns. A novel sliding cluster attention module is designed to address global feature dependency of self-attention mechanism in fMRI time series. Comprehensive experimental results demonstrate that our STCAL model can achieve competitive accuracies of 73.0 ± 4.5%, 72.0 ± 3.8%, and 72.5 ± 4.2% on the ABIDE I, ABIDE II, and ADHD-200 datasets, respectively. Moreover, the potential for feature pruning based on the co-attention scores is validated by the simulation experiment. The clinical interpretation analysis of STCAL can allow medical professionals to concentrate on the discriminative regions of interest and key time frames from fMRI data.**

***Index Terms*— Attention deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), co-attention learning, computer-aided diagnosis (CAD), discriminative localization, functional magnetic resonance imaging (fMRI).**

Manuscript received 6 January 2022; revised 16 September 2022 and 6 January 2023; accepted 31 January 2023. Date of publication 17 February 2023; date of current version 6 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62202399, Grant 61876162, Grant U21A20512, and Grant 61871272; in part by the Research Grants Council of the Hong Kong SAR under Grant PolyU11211521 and Grant PolyU15218622; in part by the Open Project of BGIShenzhen under Grant BGIRSZ20200002; and in part by the City University of Hong Kong (Dongguan). *(Corresponding author: Zhi-An Huang.)*

Rui Liu, Yao Hu, and Ka-Chun Wong are with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong, Shenzhen Research Institute, Shenzhen 518060, China (e-mail: rliu38-c@my.cityu.edu.hk; yaohu4-c@my.cityu.edu.hk; kc.w@cityu.edu.hk).

Zhi-An Huang is with the Research Office, City University of Hong Kong (Dongguan), Dongguan 523000, China, and also with the City University of Hong Kong, Shenzhen Research Institute, Shenzhen 518060, China (e-mail: huang.za@cityu.edu.cn).

Zexuan Zhu is with the National Engineering Laboratory for Big Data System Computing Technology and the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: zhuzx@szu.edu.cn).

Kay Chen Tan is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: kctan@polyu.edu.hk).

Digital Object Identifier 10.1109/TNNLS.2023.3243000

I. Introduction

NEURODEVELOPMENTAL disorders are disabilities associated primarily with functional abnormalities of the neurological system and brain [1]. As the prevalent neurodevelopmental disorders, attention deficit/hyperactivity disorder (ADHD) and autism spectrum disorder (ASD) were estimated to cause severe mental health problems in 20% of adolescents worldwide [2]. Although symptom-based criteria methods have been in practice for the clinical diagnosis of neurodevelopmental disorders, their subjective decision-making process tends to misdiagnose and overdiagnose subjects, especially for mild cases [3]. Functional magnetic resonance imaging (fMRI) can examine spontaneous brain activity noninvasively through detection of the alterations in blood-oxygen-level-dependent (BOLD) signals. Based on this neuroimaging technique, various computer-aided diagnosis (CAD) approaches have sprung up in this decade.

Recent efforts to explore neurological biomarkers have been dedicated to the functional connectivity (FC) analysis of BOLD time series between specific brain regions of interest (ROIs) [4]. The most commonly used FC measure is the Pearson correlation coefficient for the functional interplay between different brain regions. Plenty of studies have proven the feasibility of FC for the identification of brain disorders such as ASD [5], ADHD [6], and Alzheimer's disease [7]. However, FC measures always generate numerous features, with only a tiny fraction of them containing helpful classification information. Most FC-based CAD methods primarily rely on hand-crafted feature engineering to select discriminative features, which are subject to domain knowledge and modeling skills. Moreover, the symmetric FC matrices based on the pairwise correlation measures could be inapplicable to ROI-level analysis. By comparison, deep learning CAD methods can take advantage of the end-to-end framework to automatically localize the discriminative ROIs through data-driven modeling.

With the advancement of deep learning, sequence transduction models such as recurrent neural network (RNN) and long short-term memory (LSTM) have been used to achieve decent performance in time-series neuroimaging applications [8]. Under a dynamical framework evolving in time, temporal variations can be characterized to explore the role of anatomy in

the transition between important signal patterns [9]. Recently, the transformer architecture [10] emerged as one of the most popular techniques by revolutionizing the use of attention mechanisms. The major innovation of transformer and its variants (a.k.a. X-formers [11]) is to replace the recurrent units with the self-attention mechanism for computing representations of its input and output by relating different positions of a single sequence. By applying X-formers in the time-series data, key time frames in fMRI scanning can be well-deciphered based on the learned attention map for model interpretation. However, unlike common sequential data such as text, audio, and video, the fMRI data involve the restriction of voxelwise analyses to a set of ROIs. Given dozens or hundreds of defined ROIs, the multichannel nature and synchronization property add additional complexity and challenges to capturing the spatial correlation between different ROIs [12]. Therefore, it is necessary to extend the X-formers by mutually modeling the spatial and temporal dependence in fMRI data.

Co-attention learning is one of the promising solutions to address this issue. It has achieved popularity in multimodal tasks such as visual question-answering [13] and multiscale image recognition [14]. The basic rationale behind co-attention learning is to simultaneously operate multiple input modalities and jointly learn the attention weights to exploit the multimodal interactions. Co-attention learning can bridge the interdependencies from related input modalities to refine the attention representation by complementing attention generation from each other [15].

To take this advantage, we introduce co-attention learning to extend the self-attention mechanism of X-formers. Accordingly, the original fMRI signals are represented from two perspectives, i.e., the temporal-level and spatial-level features. We take the signals acquired from different ROIs at each time point as temporal-level features. In contrast, the signals collected from each ROI at different time points are regarded as spatial-level features. In this article, we propose a spatial–temporal co-attention learning (STCAL) model to diagnose mental disorders using time-series fMRI data. STCAL is characterized by two major components, namely the sliding cluster attention (SCA) module and the guided co-attention (GCA) module. Specifically, SCA is designed to extract dynamic local feature representations for temporal-level features instead of entirely relying on self-attention mechanism in a global way. GCA is developed to jointly learn spatial–temporal attention representations with the guidance of cross-modality matching. Based on SCA and GCA, the co-attention learning network (CALN) is further established to perform feature representation and fusion. Sequentially, a simple attention-aware classification network is concatenated to achieve the classification of mental disorders. Based on the experiment results of real-world datasets, STCAl achieves competitive accuracies of 73.0 $\pm$ 4.5%, 72.0 $\pm$ 3.8%, and 72.5 $\pm$ 4.2% in ABIDE I, ABIDE II, and ADHD-200, respectively via tenfold cross-validation (tenfold CV). The main contributions of this article can be summarized as follows.

1) We develop a novel SCA module to detect similar local signal patterns for aggregating local attention representations on time-series fMRI data with a sliding fusion block. This study provides a new clue to dynamically detect local key time frames on time-series fMRI data based on an effective self-attention mechanism.
2) The GCA module is proposed to achieve the attention interaction between spatial- and temporal-level features to learn the fine-grained co-attention representations. To the best of our knowledge, this is the first attempt to apply the co-attention mechanism to simultaneously model the spatial–temporal correlation of time-series fMRI data.
3) Based on SCA and GCA, the STCAL model is constructed to diagnose mental disorders from time-series fMRI data in an end-to-end manner. The learned co-attention scores enable the identification of discriminative ROIs and key time frames in fMRI data for assisting medical professionals in the personalized diagnosis and clinical decision-making.

The rest of this article is organized as follows. In Section II, we briefly review the related work regarding the attention mechanism and discriminative localization in brain MRI data. In Section III, we introduce the detail of the proposed framework. In Section IV, we evaluated the proposed STCAL and compared it with the state-of-the-art (SOTA) approaches. Moreover, the components and properties of our network are thoroughly examined. Finally, this article is concluded in Section V.

II. Related Works

In this section, we systematically review the related works of the attention mechanism in mental disorders diagnosis based on MRI data. Then, we also review the discriminative localization methods for disorders' diagnosis in the neuroimaging domain from conventional methods and deep learning methods.

A. Attention Mechanism in Neuroimaging

The attention mechanism first emerged in the deep learning community as an improvement over the encoder–decoder-based machine translation system in [16]. The intuition behind attention can be modeled as a simple weighting operation in machine learning. It helps models focus on specific parts of data to learn more discriminative patterns through different weighted assignments. With the great progress of the attention mechanism, it has become an essential component of neural networks with broad applications in natural language processing, speech processing, and computer vision [11]. Recently, the attention mechanism has also played an important role in advancing the deep learning application in neuroimaging. For example, Daza et al. [17] applied precalculated visual attention (saliency) mask on MRI images along with support vector machine (SVM) to diagnose Alzheimer's disease (AD). Olfa et al. [18] proposed a multiviewed (axial, coronal, and sagittal) attention estimation method to select salient

ROIs from structural MRI for AD classification. Different from the above two methods of estimating attention masks, Liu et al. [19] proposed an attention module based on extra-trees algorithm to select the most important ASD-related features from fMRI time-series data. Dai et al. [20] designed a dual-attention recurrent network to help the model concentrate on small but task-relevant parts of the input MRI while reducing the computational and memory costs. Broadly speaking, the attention mechanism used in the above studies can be categorized as hard attention that selects some salient parts (hard selection) from the precalculated attention scores. The hard attention mechanism in these works always acts as a separate feature selection component that is difficult to be integrated into the end-to-end diagnostic model.

In contrast, the soft attention mechanism is proposed to aggregate attention by the weighted average in a soft selection way. The soft attention makes the neural network model smooth and differentiable while easy to train through back-propagation [21]. Thanks to these merits, in recent years, increasing CAD methods have been developed based on the SOTA models of the soft attention mechanism, e.g., Transformer and BERT [11]. For instance, Niu et al. [22] applied the self-attention mechanism to identify patients with ASD by integrating multiatlases FC features. Inspired by DenseNet [23], Zhang et al. [24] introduced the attention mechanism to dense connection to reinforce visual focus in different network layers for AD classification. Except for Transformer-related works, Chen et al. [25] incorporated the attention mechanism in graph neural network (GNN) to identify important brain regions and connections contributing to the classification of ASD from structural and functional MRIs. Moreover, the soft attention mechanism has recently been incorporated with multitask learning and multimodality fusion for computed tomography image segmentation [26] and MRI-based mental disorders diagnosis [27], respectively. Lian et al. [28] also used $1 \times 1 \times 1$ convolution layer to generate attention map in a multitask framework for estimating dementia status with structural MRI. However, these existing attention-based methods are only based on visual attention or spatial attention (the attention of FC features) from MRI data. They cannot catch the spatial–temporal correlation patterns embedded in the data, which are subtle but significant for joint feature representation.

B. Discriminative Localization for Disorders Diagnosis

Detecting the discriminative regions from the brain MRI scan serves as the fundamental premise to construct a reliable classification model for diagnosing mental disorders [28]. In general, most existing CAD methods for discriminative localization can be loosely split into two categories, i.e., conventional methods and deep learning methods. The conventional methods typically take advantage of prior knowledge and experience to extract salient features for discriminative localization and mental disorders' classification. For example, Fan et al. [29] proposed a regional grouping for diagnosis of schizophrenia to segment brain MRI into multiple discriminative regions based on a preconstructed morphological profile. According to the predetermined Colin27 template in [30], Zhang et al. [31] designed a group comparison method to identify statistically significant local patterns for distinguishing AD subjects from typical controls (TCs) based on the relevant landmark locations. Bassett et al. [32] selected 90 remarkable regions from the commonly used anatomical automatic labeling (AAL) atlas in neuroimaging research to extract FC features for the identification of schizophrenia via SVM. Moreover, certain ROIs, such as the cerebellum and hippocampus, are regarded as high-risk areas with dysfunctional states in common mental diseases [33]. Focusing on these remarkable brain regions, the researchers of [34] and [35] extracted local image features for diagnosing schizophrenia and dementia, respectively. Generally, the conventional learning methods are mainly composed of two independent processes: discriminative regional detection and classifier construction. The main drawback of these conventional learning methods is that discriminative regional detection can be viewed as a filter method to filter out redundant and noisy regions based on predefined settings. Since discriminative regional detection is independent of classifier construction, the selected ROIs could be disabled to reflect the brain activity with discriminatory power.

Unlike the conventional methods, several deep learning approaches have been developed to automatically identify discriminative regions from MRI in a data-driven manner. For example, Lian et al. [36] presented a hierarchical fully convolutional network to detect salient brain locations at patch and region levels for AD diagnosis and mild cognitive impairment conversion prediction. Mao et al. [37] proposed a 4-D convolutional neural network (CNN) to directly extract the discriminative feature maps from 4-D MRI data for ADHD classification. To identify AD progression, Pena et al. [38] located significant areas of brain structural changes directly from the raw voxels with their end-to-end 3-D CNN architecture. Lian et al. [28] and [39] trained end-to-end models based on CNN to automatically generate the attention map for discriminative localization. Due to the fact that some mental disorders, such as ASD and ADHD, only cause minor structural and functional changes in the brain, the end-to-end principle is quite sensitive to the effectiveness of deep learning methods without any guidance of prior knowledge. Therefore, most SOTA CAD methods (e.g. [5], [40], [41]) require fixing the preset ROI atlas so as to maintain the decent performance. In theory, an end-to-end deep learning model can outperform the classical pipeline-based methods for complex classification tasks because of the ability to learn and self-optimize its behavior from data. To this end, in addition to the classification task, the proposed end-to-end deep learning framework aims to simultaneously identify the discriminative brain regions and key time frames in MRI data via the co-attention mechanism.

III. Methodology

As shown in Fig. 1, we propose an STCAL model to capture spatio-temporal discriminative representations from time-series fMRI data for brain disease diagnosis. We first design the CALN by the modular combination of the self-attention (SA) module, GCA module, and SCA module to learn fine-grained co-attention representations.

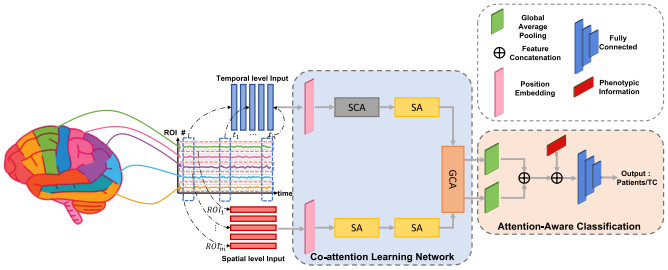


Fig. 1. Flowchart of the proposed framework.

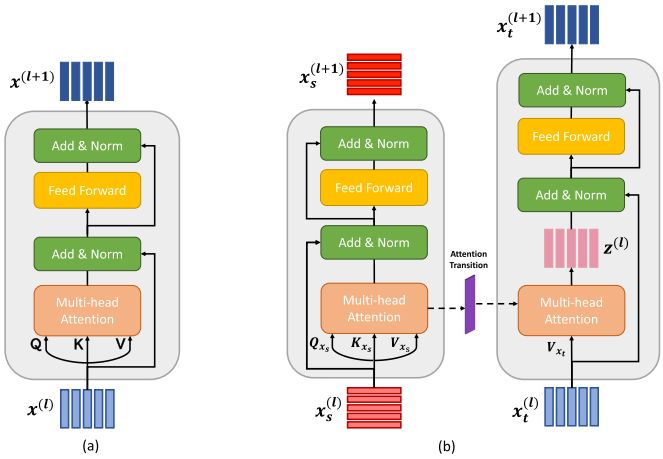


Fig. 2. Two attention modules with the multihead attention mechanism. (a) SA and (b) GCA modules are the modules located in Fig. 1.

Based on spatio-temporal co-attention representations, the attention-aware classification network is constructed to fuse spatial–temporal feature representations for classification.

A. Problem Definition

The primary goal of this work is to learn a deep co-attention model for the identification of mental disorders by exploiting spatial–temporal attention representations from MRI data. In our co-attention learning, we denote the training dataset with the sample number u as $D = \{x^i, y^i\}_{i=1}^u$, where $x_i \in X$ is the 2-D time-series fMRI input of the ith instance and $y^i \in Y$ is the corresponding training label. In this work, x_i can be represented as spatial-level feature matrices $x_s^i \in R^{m\times n}$ and temporal-level feature matrices $x_t^i \in R^{n\times m}$. Namely, $x_s^i = (x_t^i)^T$, and m and n denote the numbers of ROIs and timestamps, respectively. Given the training dataset D, the mapping function f is learned to estimate the probability for each instance in D. According to the idea of empirical risk minimization in machine learning, the optimization problem can be stated as follows:

$$f^* = \arg\min_f \sum_{i=1}^{n} \mathbb{E}_{x_s^i, x_t^i, y^i}\left[\mathcal{L}\left(f\left(x_s^i, x_t^i\right), y^i\right)\right] \tag{1}$$

where $\mathbb{E}$ is the expectation of the loss function $\mathcal{L}$ over D.

B. Preliminaries

To model the intramodal correlation of the feature representation, we introduce the SA module (the encoder layer of the transformer [10]) into our framework as a basic module. As can be seen in Fig. 2(a), the internal details of the SA module are depicted as a multihead attention block dubbed MHA followed by a pointwise feedforward layer (FFL). A residual connection is added to each sublayer along with layer normalization as the "Add & Norm" component. Three trainable matrices Q, K, and V (corresponding to *queries*, *keys*, and *values*) are computed to generate the attention block. Specifically, the dot-product similarity between *queries* and *keys* is calculated to obtain the attentional distributions on values. The resulting weighted value matrix forms the output of the attention block. By leveraging the multihead technique, multiple sets of Q, K, and V are trained to project the input embeddings into different representation subspaces to improve the representation capacity of the attention block. The output features of MHA with h heads can be formulated as

$$\mathrm{MHA}(Q, K, V) = \langle \mathrm{head}_1, \ldots, \mathrm{head}_h \rangle W^o \tag{2}$$

$$\mathrm{head}_j = \mathrm{softmax}\left(\frac{QW_j^q W_j^{k^T} K^T}{\sqrt{d}}\right) V W_j^v \tag{3}$$

where $\langle . \rangle$ is a concatenation operator, and d is the scaling factor. $[W_j^q, W_j^k, W_j^v] \in \mathbb{R}^{d_h \times d}$ and $W^o \in \mathbb{R}^{hd_h \times d}$ represent the learned linear transformation matrices for the jth head and the final head concatenation, respectively. d_h denotes the dimensionality of the output features from each head. In practice, we usually set $d_h = d/h$ to reduce the dimension of each head and keep the total computational cost the same as single-head attention with full dimensionality. Then, the FFL applies two linear projections with *gelu* activation function to the output of MHA as the following:

$$\mathrm{FFL}(x) = \mathrm{gelu}(xW_1 + b_1)W_2 + b_2 \tag{4}$$

where the linear projections are parameterized by $[W_1, W_2]$, and $[b_1, b_2]$ are the corresponding bias units. Finally, to facilitate the model optimization, two residual connections with layer normalization are, respectively, applied to the output of the MHA and FFL.

C. Co-Attention Learning Network

Based on the spatial–temporal fMRI inputs, a hybrid deep learning architecture CALN is proposed in this section to learn fine-grained co-attention representations. As shown in Fig. 1, input feature matrices are first fed to the position embedding network to encode the position information into the input features. After obtaining encoded spatial–temporal features, three-layer representation learning modules (SA, SCA, and GCA) are constructed to capture intramodal correlations, local feature importance, and intermodal correlations, respectively. The extracted co-attention representations of spatial–temporal features are then used for attention-aware classification.

1) Sliding Cluster Attention Module: The self-attention mechanism typically obtains the attention focus from the whole input features in a global way. However, local feature relationships may be ignored by the self-attention mechanism. This drawback has been verified in many recent visual task-related studies [42], [43]. To better illustrate the limitations of the standard attention mechanism used for local representation extraction, a simple example is presented in Fig. 3.

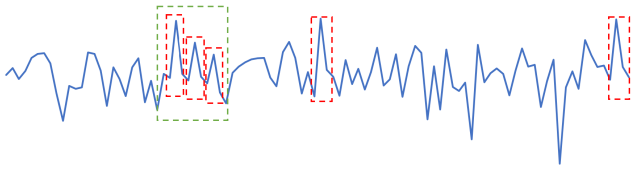


Fig. 3. Example of the temporal ROI signal pattern analysis.

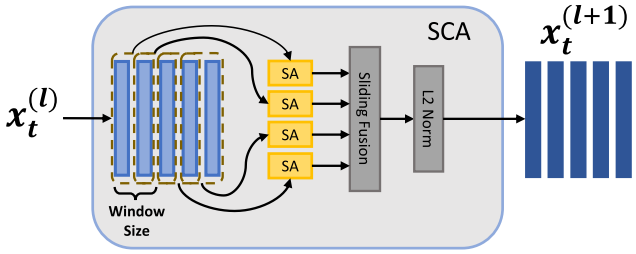


Fig. 4. Schematic of the sliding cluster attention module.

The blue polyline indicates the time-series fMRI signal from a single ROI, and the red dashed rectangles represent the tagged parts sharing similar signal patterns with different peak values. Due to the global attention computation strategy, the standard self-attention mechanism tends to focus on similar patterns with equal peak values globally while lessening attention weights of similar local patterns with different peak values (e.g., similar patterns in the green dashed rectangle). Unlike language processing tasks that use self-attention to establish global word tokens dependency, the short-time local signal yields a better estimate of BOLD responses, which is essential for time-series fMRI pattern analysis [44].

To address the above limitation, the SCA module is proposed to capture similar short-time local signals. As shown in Fig. 4, the SCA module builds independent self-attention layers based on the sliding window clusters and aggregates local attention representations with the sliding fusion block. To generate the sliding window clusters, we sample the temporal-level features' representation at the lth layer $x_t^{(l)}$ into k sequence segments $x_{t_k}^{(l)} \in \mathbb{R}^{n\times w}$ with the window size w and step size d. The number of segments k can be calculated as

$$k = \frac{n}{\lceil w - d \rceil} - 1. \tag{5}$$

Accordingly, we apply the same set of independent SA modules to the k sequence segments in parallel. Then, the sliding fusion block is used to aggregate the local attention representations from different sliding clusters through weighted summation. Finally, $l2$ normalization is applied to the aggregated attention fusion to balance the contribution of different sliding clusters. The output of the SCA module $x_t^{(l+1)}$ can be formulated as follows:

$$x_t^{(l+1)} = \frac{\left\langle \alpha^1 \mathrm{SA}\left(x_{t_1}^{(l)}\right), \ldots, \alpha^k \mathrm{SA}\left(x_{t_k}^{(l)}\right)\right\rangle}{\left\| \left\langle \alpha^1 \mathrm{SA}\left(x_{t_1}^{(l)}\right), \ldots, \alpha^k \mathrm{SA}\left(x_{t_k}^{(l)}\right)\right\rangle \right\|_2} \tag{6}$$

where α^k is the learnable weighted parameter in the sliding fusion block.

2) GCA Module: Although the self-attention mechanism can simulate intramodal dependency from temporal-level features, it fails to capture the spatial correlation between different ROIs. The spatial correlation (e.g., FC) is also informative for the diagnosis of mental disorders [12]. The co-attention mechanism facilitates the understanding of the spatial–temporal feature relationship by extracting the fine-grained representations from the joint cross-modal interactions. In this way, the spatial correlation can be effectively incorporated into the learned representations for further classification. Based on the co-attention mechanism, we propose the GCA module to learn spatio-temporal representations from pairwise input features. To model the intramodality relationship between x_s and x_t, the attention transition block (AT) is integrated into the proposed GCA module. As shown in Fig. 2(b), the GCA module is constructed in a hybrid interaction mode.

Given the intermediate spatial and temporal representations $x_s^{(l)}$ and $x_t^{(l)}$, the GCA module first computes the corresponding Q, K, and V matrices in the same way as the SA module. Since the dynamic spatial brain network involves tens of thousands of FC features for each time frame, it is rather challenging to characterize effective time-evolving brain connectivity patterns. To address this issue, the spatial attention score represented by Q_{X_s} and K_{X_s} can incorporate its own attention into the spatial representation for each time frame through our AT block. In other words, the motivation for using spatial attention scores in GCA is to guide the temporal signal extraction by concentrating on the temporal variations in those prominent ROIs (with high spatial attention scores). The spatial attention score learned from $x_s^{(l)}$ is passed to the AT block to help generate the attention score for $x_t^{(l)}$. Q_{X_s} and K_{X_s} are mainly used to allow the spatial attention calculated by the spatial SA module to flow into the temporal GA module for modeling the implicit spatio-temporal interdependencies in fMRI signal. The AT block composed of two linear transform layers aims to model the internal mapping of the attention score from $x_s^{(l)}$ to $x_t^{(l)}$. The score translation process can be formulated as follows:

$$\mathrm{AT}\left(Q_{x_s}, K_{x_s}\right) = W_{t_1}\sigma\left(W_{t_0}\left(Q_{x_s} \times K_{x_s}^T\right)\right)^T \tag{7}$$

where σ represents the activation function of ReLU, and W_{t_0} and W_{t_1} are the trainable parameters for the first and second dense layers, respectively. As such, the AT block outputs a mapping matrix $\in \mathbb{R}^{m\times m}$ to represent the spatial dependency on the temporal dynamic variations. Consequently, the MHA block of $x_t^{(l)}$ produces attention-pooled feature $z^{(l)}$ in coordination with the attention transition block of $x_s^{(l)}$. According to (2) and (3), the spatial-guided temporal representation $z^{(l)}$ can be reached as follows:

$$z^{(l)} = \mathrm{MHA}\left(Q_{x_s}, K_{x_s}, V_{x_t}\right) = \langle \mathrm{head}_1, \ldots, \mathrm{head}_h \rangle W^o \tag{8}$$

$$\mathrm{head}_j = \mathrm{softmax}\left(\frac{\mathrm{AT}\left(Q_{x_s}W_j^q, K_{x_s}W_j^k\right)}{\sqrt{d}}\right) V_{x_t} W_j^v. \tag{9}$$

3) Architecture: The modules of SA, SCA, and GCA are integrated to construct CALN. In our implementation, the pairwise spatial–temporal input $[x_s, x_t]$ is first fed to the learned

Algorithm 1 Pseudocode of the STCAL

Require: Preprocessed data X_S, X_T, phenotypic data X_P, and label Y

Ensure: Predicted probabilities of testing set Y^{te}_{pred}

$[X^{tr}_S, X^{te}_S, X^{tr}_T, X^{te}_T, X^{tr}_P, X^{te}_P, Y^{tr}, Y^{te}] \leftarrow$ Split(X_S, X_T, X_P, Y)
Initialize STCAL.
for $e = 1, \ldots, epochs$ **do** //e: # of training epoch
  // Co-attention Learning Network
  $[X^{tr'}_S, X^{tr'}_T]$ = PositionEmbedding $([X^{tr}_S, X^{tr}_T])$
  for $l = 1, \ldots, L$ **do** // l: # of attention layers
    if $l == 1$ **then**
      $X^{tr^{(1)}}_S \leftarrow$ SA $(X^{tr'}_S)$; $X^{tr^{(1)}}_T \leftarrow$ SCA$(X^{tr'}_T)$
    else if $l = 2, \ldots, L-1$ **then**
      $X^{tr^{(l)}}_S \leftarrow$ SA $(X^{tr^{(l-1)}}_S)$; $X^{tr^{(l)}}_T \leftarrow$ SA $(X^{tr^{(l-1)}}_T)$
    else
      $[X^{tr^{(L)}}_S, X^{tr^{(L)}}_T] \leftarrow$ GCA$(X^{tr^{(L-1)}}_S, X^{tr^{(L-1)}}_T)$
  end for
  // Attention-Aware Classification Network
  $\hat{X^{tr}_S} \leftarrow$GAP$(X^{tr^{(L)}}_S)$; $\hat{X^{tr}_T} \leftarrow$ GAP$(X^{tr^{(L)}}_T)$
  $X^{tr}_R \leftarrow$ Concatenate$(\hat{X^{tr}_S}, \hat{X^{tr}_T}, X^{tr}_P)$
  $output \leftarrow$ FullyConnectNetworks(X^{tr}_R)
  $Loss \leftarrow$ BinaryCrossEntropy$(output, Y^{tr})$
  STCAL.update$(Loss)$
end for
$Y^{te}_{pred} \leftarrow$ STCAL.predict $(X^{te}_S, X^{te}_T, X^{te}_P)$

position embedding network to incorporate the corresponding position information. The basic principle is to add the position sequence encodings learned from the mapping function to the input features. Afterward, SCA is used to detect the important local pattern for modeling the dynamical correlation between different timestamps. Since the sliding window approach cannot be applied to the sequential input of x_s, the SA module is adopted instead. At the end of CALN, we use the GCA module to generate the higher level attention scores of x_s and x_t based on a co-attention mechanism. Considering there could be other potential modes for the combination of SA, GCA, and SCA, we further explored the performance of other structural variants in Section IV. Based on the learned attention matrices of x_s and x_t, CALN can simultaneously locate the discriminative ROIs and time frames.

D. Attention-Aware Classification Network

Based on the pairwise outputs of CALN, a co-attention-aware classification block is further designed to diagnose multiple mental disorders. We first adopt the global average pooling (GAP) to downsample the feature map of learned spatial representation x'_s and temporal representation x'_t. Given the spatial invariance characteristic of GAP [45], we can maintain feature relationships between x'_s and x'_t while shrinking the size of neuron clusters in feature maps. The nonlinear transformation of GAP also reduces the risks of overfitting. After that, we compute the overall co-attention representation x_r as

$$x_r = \langle \hat{x_s}, \hat{x_t}, x_p \rangle \tag{10}$$

TABLE I
PHENOTYPIC INFORMATION IN ABIDE I, ABIDE II, AND ADHD-200

Dataset	Pipeline	Label	Age	Gender	Handedness	IQ
ABIDE I	C-PAC	TC	16.8±7.4	435/95	492/3/36	110.8±12.4
		ASD	17.1±8.5	443/62	455/3/47	104.8±16.9
ABIDE II	DPARSF	TC	15.0±9.3	412/180	537/29/28	113.5±12.9
		ASD	14.8±9.0	444/77	443/37/41	106.1±16.4
ADHD-200	Athena	TC	12.2±3.5	304/277	528/2/51	112.9±13.2
		ADHD	11.7±3.0	282/76	303/1/54	105.7±13.0

Gender: male/female; Handedness: right/ambidextrous/left

where x_p is the phenotypic information of subjects. Finally, a neural network built of three fully connected layers accepts x_r to make a classification of mental disorders. To ease the understanding of the proposed STCAL, the pseudocode is shown in Algorithm 1.

IV. EXPERIMENTS AND RESULTS

In this section, we first introduce the datasets and preprocessing pipelines used in this work. The details of model settings and evaluation metrics are provided. Then, we compare our STCAL with several SOTA methods. After that, ablation studies are also performed to validate the effectiveness of major components in STCAL. The performance of structural variants in GCA and CALN is also investigated in this section. Moreover, the potential of practical applications of STCAL is investigated. Finally, we visualize the learned attention maps at both the spatial and temporal levels for model interpretation.

1) Slice timing correction and realignment.
2) Motion correction and spatial normalization.
3) Frequency excursion by the bandpass filter.
4) Normalization by the MNI template.
5) Spatial smoothing using a 6-mm FWHM Gaussian filter.

A. Data Acquisition and Processing

Three public MRI datasets studied in this work are downloaded from the Autism Brain Imaging Data Exchange (ABIDE) I,[1] ABIDE II,[2] and the ADHD-200 Consortium (ADHD-200).[3] Each sample in these datasets contains functional MRI data along with phenotypic information, such as age, gender, handedness, and IQ. In this work, we only used four types of phenotypic information (i.e., age, gender, handedness, and full IQ) as shown in Table I. Among the four phenotypic information used in this work, gender and age information was available for all the samples. The data missing rates of handedness/full IQ are 28.9%/6.3%, 2.1%/8.9%, and 0.9%/8.6% in ABIDE I, ABIDE II, and ADHD-200 datasets, respectively. We referred to the previous literature to handle the missing values of handedness and full IQ. For handedness, we simply assigned "right" to those samples without handedness values because most people are right-handed. According to the previous work [46], the mean IQ over all the valid samples was an alternative to the missing IQ values.

[1]https://fcon_1000.projects.nitrc.org/indi/abide
[2]https://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html
[3]https://fcon_1000.projects.nitrc.org/indi/adhd200/

ABIDE I aggregated 1035 valid samples from 17 international imaging sites, where 505 are individuals with ASD and 530 are TCs. The ABIDE II dataset consists of 19 different sites with 1113 valid subjects, including 521 ASD participants and 592 TCs. The ADHD-200 dataset is also an aggregated dataset with 939 valid samples from a collaboration of eight international imaging sites. The ADHD-200 dataset contains 358 ADHD subjects and 581 TCs. Due to the data-sharing efforts of ABIDE I and ADHD-200, the preprocessed data can be downloaded according to the Configurable Pipeline for the Analysis of Connectomes (C-PAC)[4] of ABIDE I and Athena[5] of ADHD-200, respectively. The ABIDE II dataset is preprocessed using the Data Processing Assistant for Resting-State fMRI (DPARSF)[6] pipeline. Although the implementation details of these three pipelines are slightly different, the main processes can be summarized as the following steps:

In this study, we apply the Craddock 200 (CC200) functional parcellation atlas of the brain to generate preprocessed mean time-series matrices with 190 ROIs. Since ABIDE I, ABIDE II, and ADHD-200 are multisite datasets where the length of time course varies from site to site, we adopt a random cropping approach to reduce overfitting and coordinate the same sequence length as inputs for regulation, which was demonstrated by the previous work [37]. To guarantee the fairness of simulation experiments, we totally follow their proposed procedure to split the whole dataset into the training and testing sets prior to performing cropping operations. Therefore, there is no sequence overlap between the split training set and testing set. In this work, each sample is randomly cropped into ten sequences of length T from time-series fMRI data. We set $T = 90$, according to the length of the shortest time series. Therefore, each cropped sample is regularized to 90×190, i.e., sequence length $\times$ number of ROIs. To evaluate the proposed model, as shown in Fig. 5, 10-fold cross-validation with validation set and testing set is performed to split the total dataset into k sets ($k = 10$), where a set is selected as the testing set one by one, whereas one of the remaining sets is used as the validation set and the other $k - 2$ sets are used as the training sets until each set has been evaluated as a testing set. In this setting, the training set is used for model fitting and the validation set is used to evaluate the iterative model based on key metrics in the validation set and select the best one according to the model checkpoints during training. Finally, we call back the saved model with the best trainable weights and evaluate it on the testing set, which was unseen in the model before. Moreover, since the ADHD-200 dataset was officially split into the training/testing sets (IS) for global competition, we also evaluated the performance of the proposed model in this IS scheme.

B. Experimental Settings

The model parameter settings of STCAL are shown in Table II. For a more comprehensive evaluation, the performance of our model is assessed by four common metrics, namely, accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AUC). These metrics are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{11}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{13}$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative values, respectively. According to all possible SEN and 1-SPE pairs, the receiver operating characteristic (ROC) curve is plotted by varying the threshold performed on the classification score. By computing the area under the ROC curve, AUC gives an aggregated measure of performance overall potential classification thresholds. In general, the prediction models with higher AUC scores can better distinguish patients from TCs. In this work, the final predicted label for each subject is determined by the majority voting scheme.

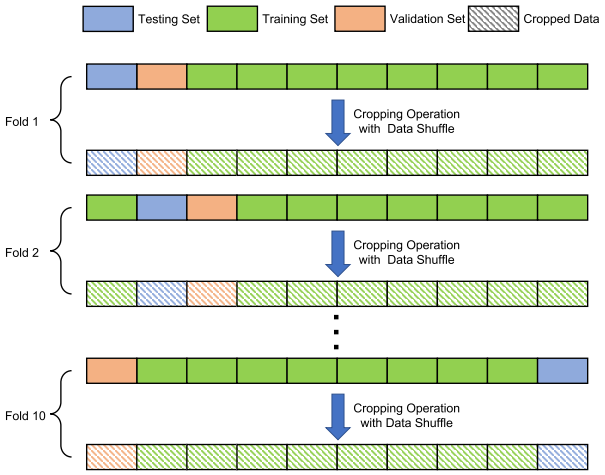


Fig. 5. Data splitting and augmentation scheme.

TABLE II
MODEL PARAMETER SETTINGS OF STCAL

# of cluster in SCA	5	Window size of SCA	30
# of head in GCA	10	Max position embeddings	512
Dropout rate	0.1	Learning rate	10^{-5}
Batch size	128	Epochs	30
Configuration of fully connected networks		64-Dropout(0.5)-10-1 (Three layers)	
Activation function to attention layers		Gaussian error linear units (gelu)	
Activation function to output layers		Sigmoid	

C. Comparison With the SOTA Methods

In this section, the STCAL is compared with the SOTA methods on ABIDE I, ABIDE II, and ADHD-200 in terms of accuracy, respectively, as shown in Table III. Due to the different adopted sampling criteria (e.g., IQ-matched participants, similar MRI acquisition protocols, and site-matched participants), a portion of previous works are subject to relatively

[4]https://preprocessed-connectomes-project.org/abide/
[5]https://www.nitrc.org/projects/neurobureau
[6]https://rfmri.org/DPARSF/

TABLE III
PERFORMANCE COMPARISON ON ABIDE I, ABIDE II, AND ADHD-200

Model	Classifier	Validation	Sample #	phenotypic	Accuracy(STD)
For ABIDE I (ASD)					
Ours	STCAL	10-fold CV	1035	A/G/H/IQ	**73.0%(4.5%)**
Ours w/o pheno	STCAL	10-fold CV	1035	N.A.	72.1%(4.8%)
Chen 2022 [25]	GAN	10-fold CV	1007	A/G	70.8%(2.9%)
Amuqhim 2021 [47]	SAE	10-fold CV	1035	N.A.	70.8%(N.A.)
Zeinab 2020 [48]	CNN	10-fold CV	1035	A/G	70.2%(N.A.)
Huang 2020 [5]	DBN	10-fold CV	1035	A/G/H/IQ/E	72.5%(3.7%)
Eslami 2019 [49]	AE	10-fold CV	1035	N.A.	70.3%(N.A.)
Heinsfeld 2018 [50]	AE+ANN	10-fold CV	1035	A/G/H	70.0%(N.A.)
For ABIDE II (ASD)					
Ours	STCAL	10-fold CV	1113	A/G/H/IQ	**72.0%(3.8%)**
Ours w/o pheno	STCAL	10-fold CV	1113	N.A.	71.4%(4.1%)
Liu 2021 [51]	BL	10-fold CV	1043	N.A.	65.3%(N.A.)
Chen 2020 [52]	SVM	10-fold CV	250	A/G	65.0%(N.A.)
Aghdam 2019 [53]	CNN	10-fold CV	343	N.A.	70.0%(N.A.)
Zhao 2019 [54]	CAE	IS	693	N.A.	65.3%(N.A.)
For ADHD-200 (ADHD)					
Ours	STCAL	10-fold CV	939	A/G/H/IQ	72.5%(4.2%)
Ours w/o pheno	STCAL	10-fold CV	939	N.A.	71.5%(3.9%)
Ours	STCAL	IS	939	A/G/H/IQ	**74.3%**(N.A.)
Dou 2020 [6]	EM-MI	IS	782	N.A.	70.2%(N.A.)
Mao 2019 [37]	4D-CNN	IS	788	N.A.	71.3%(N.A.)
Zou 2017 [55]	3D-CNN	IS	730	A/G/H/IQ	69.2%(N.A.)
Tan 2017 [56]	Linaer SVM	10-fold CV	217	A/G/H/IQ	68.6%(1.7%)
Ghiassian 2016 [56]	RBF-SVM	IS	940	A/G/H/IQ/S	70.0%(N.A.)

Note: Deep Belief Network (DBN), Autoencoder (AE), Sparse Autoencoder (SAE), Radial Basis Function (RBF), Convolutional Autoencoder(CAE), Broad Learning (BL), Graph Attention Network (GAN). We abbreviate age, gender, handedness, eye status, and site information as A, G, H, E, and S, respectively.

small sets of data for their model evaluation. Therefore, their model performance could significantly decline when using the whole dataset. To evaluate the generalization ability and reliability of the proposed model, the tenfold CV is conducted on the whole datasets of ABIDE I, ABDIE II, and ADHD-200, respectively.

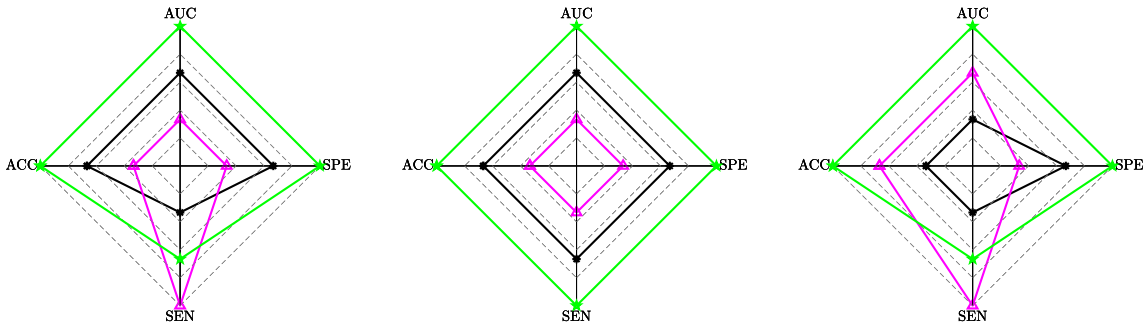
In comparison between STCAL and other SOTA models, STCAL obtains a competitive classification performance of 73.0% (SEN: 79.8%, SPE: 65.9%, AUC: 78.2%), 72.0% (SEN: 74.4%, SPE: 69.3%, AUC: 76.8%), and 72.5% (SEN: 79.2%, SPE: 64.6%, AUC: 78.0%) on ABIDE I, ABIDE II, and ADHD-200, respectively. Considering the natural data split of ADHD-200 for global competition, the proposed model is also evaluated via IS validation, where the validation set accounts for 20% of the training set. As a result, STCAL achieves a robust accuracy of 74.3% (SEN: 88.9%, SPE: 54.2%, and AUC: 83.8%) using the IS of ADHD-200. The comparison results show that most deep-learning-based methods outperform the conventional machine-learning-based methods on these datasets. In particular, the DBN-based model in [5] and CNN-based models in [53] and [37] achieve the highest accuracy reported to date w.r.t. the whole datasets of ABIDE I, ABIDE II, and ADHD-200, respectively. According to the results reported in Huang's work [5], the automated hyperparameter tuning method based on Bayesian optimization (BO) was demonstrated to further improve the classification performance at the expense of additional computational resources and elapsed time. We also evaluate the effect of BO-based hyperparameter tuning on STCAL with

TABLE IV
PARAMETER RANGE AND OPTIMUM VALUES OF BO IN STCAL

Parameters	Search Range	Optimal Value
Batch Size	[16, 128]	114
Dropout Rate	[0.1, 0.5]	0.41
Learning Rate	[1, 0.01]	0.81
Unit # of Fully Connected Layer 1	[32, 128]	112
Unit # of Fully Connected Layer 2	[10, 32]	14

30 optimization iterations, the same as that in Huang's work [5]. As a result, STCAL achieves the mean accuracy of 74.8 $\pm$ 3.7%, with an increase of 1.8% over the baseline (without BO). This result demonstrates that STCAl can also benefit from BO-based hyperparameter tuning to obtain a competitive performance comparable to that reported in Huang's work, i.e., the mean accuracy of 74.5 $\pm$ 3.1% in the ABIDE I dataset. The parameter range and optimum values are listed in Table IV. Furthermore, the recently popular GNN-based model [25] achieves 70.8% accuracy on ASD classification (use fMRI features only), which is still 2.2% less accurate than our model. Recent evidence suggests [57] that phenotypic information can boost classification accuracy significantly, and we performed separate experiments on the ABIDE I, ABIDE II, and ADHD-200 datasets without using phenotypic information. As can be seen from Table III, our proposed STCAL without using phenotypic information still achieves robust classification performance. Specifically, diagnostic accuracy drops slightly by 0.9%, 0.6%, and 1.0% on the ABIDE I, ABIDE II, and ADHD-200 datasets, respectively, but still outperforms the compared work that did not use phenotypic information. By removing phenotypic information, such an extent of performance degradation is also observed in the previous deep learning work for MRI diagnosis [27] and therefore could be reasonably acceptable. Thanks to the effective co-attention learning for spatial–temporal representation of MRI data, the proposed framework achieves supreme accuracy among the compared SOTA methods.

As the first co-attention network applied in neuroimaging analysis, STCAL is also compared with two SOTA co-attention networks for performance comparison: VliBERT [15] and MCAN [13]. Specifically, VliBERT extends the BERT architecture to a two-stream model based on a co-attention module that directly exchanges the K and V matrices of two streams. MCAN introduces a modular co-attention layer cascaded to a deep neural network. Since VliBERT and MCAN are originally proposed for the vision-and-language tasks, their entire network structure is incompatible with the neuroimaging diagnostic task. Therefore, we replace the co-attention learning module GCA in STCAL with their core co-attention modules. The classification results in tenfold CV are summarized in Fig. 6. The overall performance of STCAL still exceeds VilBERT and MCAN by dominating 75% (9/12) of evaluation metrics in this experiment. The experimental results also demonstrate that STCAL can better coordinate the tradeoff between specificity and



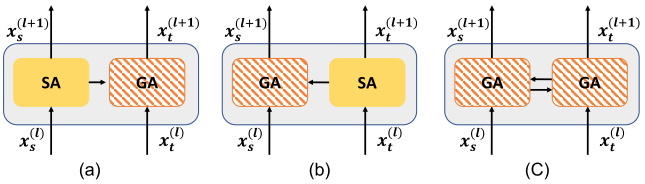
Method	Mode	ABIDE I				ABIDE II				ADHD-200			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
VilBERT	Stacking	70.9%	77.1%	63.1%	76.9%	70.7%	**75.1%**	65.7%	76.2%	70.1%	78.1%	57.1%	74.9%
MCAN	Encoder-Decoder	70.2%	**80.2%**	59.6%	76.3%	70.6%	74.1%	66.7%	75.9%	70.4%	**80.3%**	55.8%	75.4%
STCAL	Fusion	**73.0%**†	79.8%	**65.9%**†	**78.2%**	**72.0%**	74.4%	**69.3%**†	**76.8%**	**72.5%**†	79.2%	**64.6%**†	**78.0%**†

The † marked results are significantly better (Wilcoxon rank-sum test with p-value correction, α = 0.05). On each evaluation matric, the three methods are ranked from 1 to 3.

Fig. 6. Performance comparison between STCAL and other co-attention model.

TABLE V
PERFORMANCE COMPARISON BETWEEN GCA VARIANTS

Method	ABIDE I				ABIDE II				ADHD-200			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
$SA \rightarrow GA$	**71.8%**	77.3%	**66.6%**	**79.0%**	**71.0%**	**74.6%**	**67.5%**	**76.5%**	**71.3%**	76.9%	**60.1%**	**76.5%**
$GA \leftarrow SA$	70.7%	**79.8%**	54.6%	76.0%	70.2%	73.0%	67.1%	76.2%	70.9%	79.9%	56.9%	75.9%
$GA \leftrightarrows GA$	71.2%	78.2%	63.8%	78.0%	70.0%	73.9%	65.7%	75.9%	70.9%	**80.4%**	56.3%	76.0%



To better elaborate on the structure of GCA variants, we denote the right half of the GCA as guided attention (GA).

Fig. 7. Variants of the GCA module. (a) SA→GA (GCA). (b) GA→SA. (c) GA⇆GA.

sensitivity, thus reaching the highest AUC values on these three datasets.

D. Structure Exploration

1) Effectiveness of Different GCA Variants: To investigate the potential of GCA-related structures, we also evaluate the performance of the other two GCA variants, whose architectures are shown in Fig. 7. The major difference between GCA variants lies in how the guided attention flows between the spatial and temporal input modalities. That is, the key to the experiment is to explore which one serves as the target modality and can benefit from the other one as the source modality with attention scores. The SA → GA module in Fig. 7(a) represents the baseline of the proposed GCA module, which uses the spatial feature attention score to guide the generation of temporal feature representation. In contrast, the GA ← SA module in Fig. 7(b) uses the temporal feature attention score to guide the generation of spatial feature representation through the attention transition unit. As shown in Fig. 7(c), the GA ⇆ GA module fully takes advantage of the learned attention score from each input stream to generate the feature representation mutually. In this scenario, the GA ⇆ GA module provides sufficient co-attention interactions between spatial- and temporal-level features. These three modules are assembled separately in the same setting using a three-layer deep learning architecture. To be fair, the SCA module is not introduced in this part. As shown in Table V, 83.3% (10/12) of evaluation metrics are dominated by SA → GA on three datasets. The SA → GA module can characterize the effective time-evolving brain connectivity patterns by concentrating on the temporal variations in those prominent ROIs. In contrast, uncontrolled confounding factors undermine the temporal dependency on the spatial dynamic variations because each MRI dataset was sampled in different conditions, e.g., participant, scan time, and interval time. Due to the time-varying nature of the fMRI signal, the contribution of temporal attention score is relatively limited to spatial representational learning. This result is in agreement with the discovery of the recent works [58] that the performance of fMRI classification can benefit from modeling the temporal representation with the guidance of spatial dependency, i.e., the learned attention score from spatial-level features. Therefore, as we expect, the overall performance of the baseline SA → GA module succeeds GA ← SA and GA ⇆ GA modules

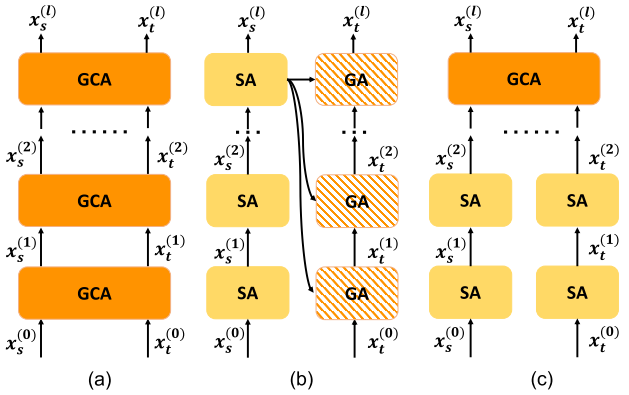


Fig. 8. Variants of deep co-attention learning networks. (a) Stacking mode. (b) Encoder–decoder mode. (c) Fusion mode.

on all the evaluated datasets. This result demonstrates that the GCA module can well model the spatial–temporal representation from fMRI data by a decent co-attention learning mechanism.

2) Effectiveness of Different CALN Variants: Similarly, the proposed CALN can be extended to different representative modes, stacking mode (CALN_{sm}), encoder–decoder (CALN_{ed}), and fusion mode with GCA (CALN_{fm}, as baseline). The architectures of these variants are illustrated in Fig. 8. The compared stacking mode and encoder–decoder mode are adopted to perform co-attention learning by VilBERT and MCAN, respectively. Therefore, we further explored how these construction modes perform with the possible combinations of the proposed GCA and SA modules. Also, to be fair, the SCA module is not used in this part. The stacking mode simply stacks the GCA modules in a three-layer deep learning architecture by hierarchically refining the co-attention representations. The encoder–decoder mode takes three stacked SA modules as an encoder to transform the spatial features into a fixed-length internal embedding. As a decoder, the integrated GA modules map the internal embedding to output the spatial-guided temporal features in a recurrent way. In the baseline setting, two SA module pairs compose the bottom network in parallel over the input. The GCA module acts as a tower network to conduct co-attention learning based on the high-level spatial–temporal representation. The experimental results for performance evaluation are tabulated in Table VI. As expected, CALN_{fm} outperforms the other two variants in terms of 75% (9/12) evaluation metrics. Specifically, the stacking mode injects spatial attention from low to high levels into the temporal attention representation by directly stacking the GCA modules. The performance degradation of CALN_{sm} could be attributed to the fact that the low-level spatial feature extraction by SA (within GCA) cannot guarantee to provide positive guidance to the temporal feature representation. This finding is consistent with the experimental results reported in the article of MCAN [13], which also compared the encoder–decoder mode with the stacking mode to address the visual question-answering problem. It is demonstrated that since the learned self-attention from bottom SA modules is inaccurate compared with that from the top SA modules, the word dependencies learned by bottom SA modules fail to improve the co-attention representation learning for images.

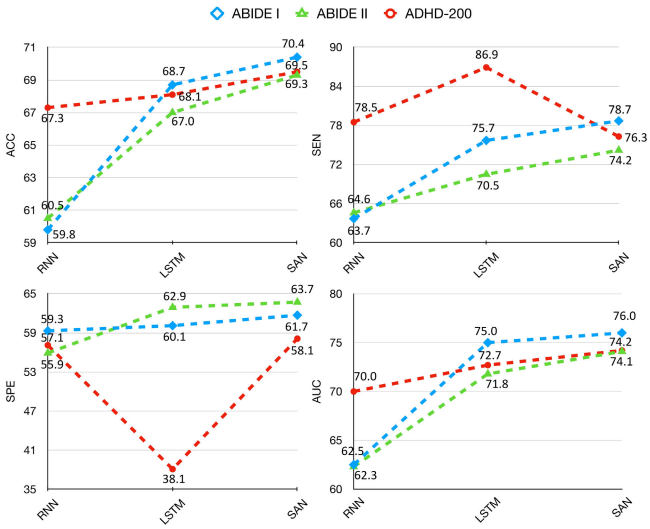


Fig. 9. Performance comparison between RNN, LSTM, and SAN.

As for CALN_{ed}, the main idea is to exploit the dependence of high-level spatial features to hierarchical temporal features (from low order to high order) for spatio-temporal co-attention learning while limiting the amount of low-level spatial information flow into the temporal GA modules. However, the global spatial dependence represented by high-level spatial attention could be detrimental to modeling low-order spatio-temporal dynamics with bottom GA modules. Therefore, the high-level spatial attention cannot be generalized effectively to the co-attention learning for hierarchical temporal features. As a result, CALN_{fm} achieves the best overall performance on three multisite fMRI datasets. Moreover, all the three variants show reliable results in this group of experiments, which demonstrates the feasibility of the co-attention mechanism in dealing with dual-modality fMRI data. Due to the lack of the SCA module, CALN_{fm} suffers a slight performance degradation compared with the best mode of the STCAL model.

E. Ablation Study

1) Effectiveness of Basic Self-Attention Network: As discussed above, the SA module is leveraged as a basic building block to develop the proposed SCA module and GCA module. Here, we perform an experiment to assess the effectiveness of the SA-based network (SAN) in our framework by comparing the other two sequence transduction models, i.e., RNN and LSTM. Since they cannot be directly applied for co-attention learning, in this case, the input data are not transformed into the spatial and temporal level features. It needs to be noted that the same settings of network configurations are adopted for a fair comparison. The corresponding results of performance comparison are presented in Fig. 9. SAN achieves the supreme comprehensive performance dominating 91.7% (11/12) of evaluation metrics, which shows the effectiveness of SAN in processing variable-length sequence data. SAN also obtains the best AUC score on those three datasets. It indicates that the SAN is capable of learning unbiased representations and performing reliable diagnostic tasks. Unlike

TABLE VI
PERFORMANCE COMPARISON BETWEEN CALN VARIANTS

Method	ABIDE I				ABIDE II				ADHD-200			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
CALN_{sm}	70.7%	72.6%	68.3%	77.2%	70.5%	74.2%	64.7%	75.8%	70.0%	76.9%	58.9%	75.6%
CALN_{ed}	70.6%	72.1%	**68.8%**	77.1%	70.4%	**74.8%**	65.7%	75.8%	70.2%	**77.9%**	57.8%	75.9%
CALN_{fm}	**71.8%**	**77.3%**	66.6%	**79.0%**	**71.0%**	74.6%	**67.5%**	**76.0%**	**71.3%**	76.9%	**60.1%**	**76.5%**

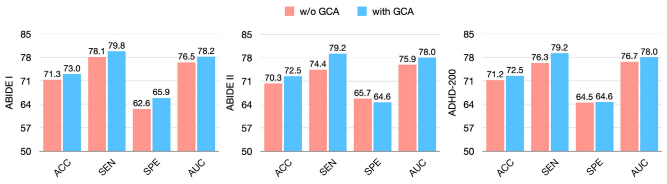


Fig. 10. Performance comparison to evaluate the effectiveness of GCA.

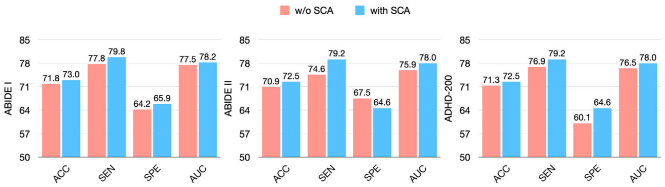


Fig. 11. Performance comparison to evaluate the effectiveness of SCA.

RNN and LSTM, SAN leverages the self-attention mechanisms to capture long-range mutual dependencies to obtain global representations. Therefore, SAN is competent in the extraction of the correlation patterns of time-series fMRI features and obtains better diagnostic performance. Moreover, LSTM slightly outperforms RNN in terms of ACC and AUC. However, LSTM fails to strike a balance between sensitivity and specificity, especially in the ADHD-200 dataset.

2) Effectiveness of GCA Module: Based on the SA modules, the proposed GCA module is a key component of STCAL to achieve the fine-grained co-attention representation between spatial- and temporal-level features. To evaluate its effectiveness, we design another version of the framework as a baseline (i.e., STCAL w/o GCA) by replacing the GCA module with two SA modules. As can be seen from Fig. 10, compared with STCAL w/o GCA, 91.7% (11/12) of the evaluation indicators are dominated by the proposed STCAL. This is because the GCA module excels at bridging interdependencies from spatial and temporal input modalities, which promotes the fine-grained representation learning from fMRI data. Compared with the SOTA methods, STCAL w/o GCA still shows overall competitive performance. Especially for the ADHD-200 dataset, STCALF w/o GCA outperforms the best model reported to date (e.g., Mao et al. [37]) in terms of accuracy. This result reveals that even without co-attention learning, the spatial-level features of fMRI can also provide effective complementary information for the classification of mental disorders.

3) Effectiveness of Sliding Cluster Attention Module: For capturing the local temporal patterns in the time-series fMRI data, the SCA module is another key component of STCAL developed by the SA modules. We also design a baseline framework by likewise substituting the SA modules for the SCA module (denoted as STCAL w/o SCA). Fig. 11 shows the results of performance comparison between STCAL and STCAL w/o SCA. As expected, the involvement of SCA can improve the overall performance of STCAL on these three datasets.

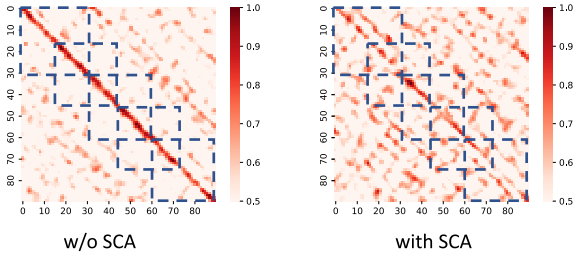


Fig. 12. Effects of gender and age are evaluated on prediction accuracy as confounding factors. The sample numbers are given in brackets.

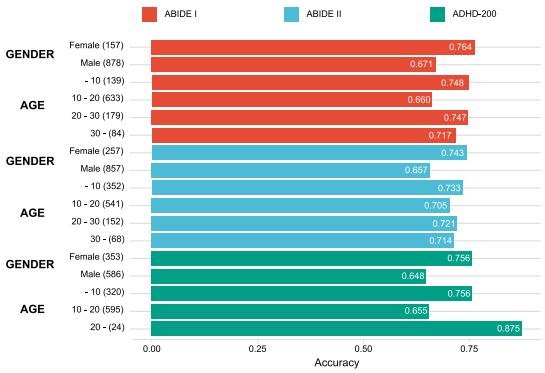


Fig. 13. Attention score matrices in the first attention layer of STCAL w/o and with SCA module on the ABIDE I dataset.

To illustrate the key to success in SCA, the attention score matrices in the first layer of STCAL are retrieved and normalized for visualization analysis, as shown in Fig. 12. It is shown that with the aid of SCA, STCAL focuses on more details of time correlation throughout the whole time series. Especially within the same range of the sliding window (blue dashed square), SCA helps the model to detect more subtle but relevant local signal patterns. These interesting observations also provide us a clue to probe the possibility of differentiating the key time frame from the whole time series.

4) Effects of Different Phenotypic Categories on Classification Accuracy: To explore how age/gender information specifically affects the diagnostic performance of STCAL on

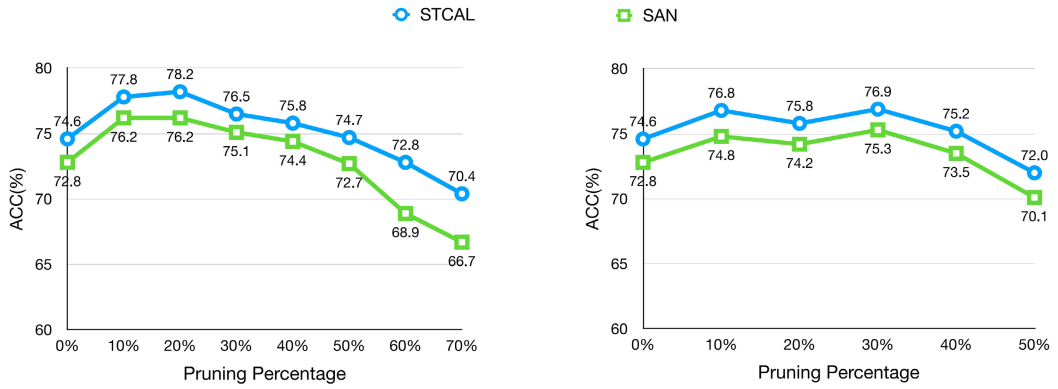


Fig. 14. Performance comparison on different pruning percentages. (a) Spatial level pruning of ROIs based on the learned attention score. (b) Temporal level pruning of time frames based on the learned attention score.

these disorders. According to the experiment setup in [5], we viewed gender/age as confounding factors and evaluated their effects on classification accuracy by regarding the specific criteria sample group (e.g., age $<$ 10) as the testing set and the remaining samples as the training set. The experimental results are presented in Fig. 13, and the sample numbers of testing sets concerning the given scenarios are indicated in brackets. As can be seen, generally speaking, there is no significant difference between accuracy and gender/age on the three datasets. Despite the class imbalance issue inevitably affecting the performance, the classification accuracy can consistently exceed 65% for each scenario. Specifically, since the male group accounts for the majority of samples in the three datasets, it is difficult to draw a conclusion about whether gender has a specific effect on the accuracy of these disorders. This result demonstrates that the performance of STCAL could be competent for particular research purposes that need to specify the recruitment strategies.

F. Potential for Practical Application

The co-attention mechanism enables our framework to locate the discriminative key ROIs and time frames from fMRI data based on the learned attention scores. In this case, we conduct simulation experiments to evaluate the potential capability of key ROIs and time frames identified by STCAL for feature-based pruning. In this section, the samples of NYU site in ABIDE I and ABIDE II are leveraged as the training set and testing set, respectively. The initial STCAL is trained on the training dataset to learn the spatial–temporal attention scores. Then we sum up the attention scores of each sample for ranking the seminal ROIs and time frames. The irrelevant ROIs and time frames with low scores are pruned in ascending order. Accordingly, the simulation experiments are conducted to evaluate the effects of feature-based pruning in either of these two ways, i.e., spatial-level pruning of ROIs and temporal-level pruning of time frames.

For a more comprehensive evaluation, the performance change in both STCAL and SAN is observed by varying the percentages of feature-based pruning. As shown in Fig. 14, feature-based pruning has a remarkably positive effect on improving the effectiveness of STCAL. As can be seen from Fig. 14, in the scenario of 20% ROIs removed, the STCAL and SAN models achieve the best accuracies of 78.2% and 76.2%, respectively. Similarly, the supreme accuracies of 76.9% and 75.3% are, respectively, achieved by STCAL and SAN in the scenario of 30% time frames removed. It is interesting to note that 50% of the pruning percentages do not cause significant performance degradation of STCAL and SAN, which still maintain over 70% of accuracy. The experimental findings further suggest that the feature-based pruning technique can improve the diagnostic accuracy while conserving computational resources. By filtering out 50% features, the pruned STCAL significantly reduces the model training time, trainable parameters, and model size by 59.8% (28.1 s/68.3 s), 68.1% (1.17 M/3.67 M), and 66.9% (14.6 MB/44.8 MB), respectively. Therefore, it is demonstrated that the discriminative localization mechanism of key ROIs and time frames in STCAL has great potential for the application of feature-based pruning.

G. Clinical Interpretation Analysis

Based on the co-attention mechanism, the proposed model can further automatically locate the remarkable ROIs and time frames on time-series fMRI data to explore the neurological patterns associated with specific mental disorders. As shown in Fig. 15, we design a visualization system based on the learned attention score matrices for auxiliary application of interpretation. For a selected input subject, the well-trained STCAL can simultaneously estimate the attention scores of key ROIs and time frames. We can preset the top-N of the most discriminative brain regions for any time frame score defined by users. Finally, the remarkably similar signal patterns and ROIs localization are correspondingly visualized in the system by dragging the time-score scroll bar. The visualization system provides medical professionals with a support tool to reduce medical errors and help coordinate follow-up.

In addition to detecting prominent individual brain regions within a short period, STCAL can also characterize global temporal dependency to identify disease-specific ROIs by aggregating the calculated attention scores of each subject

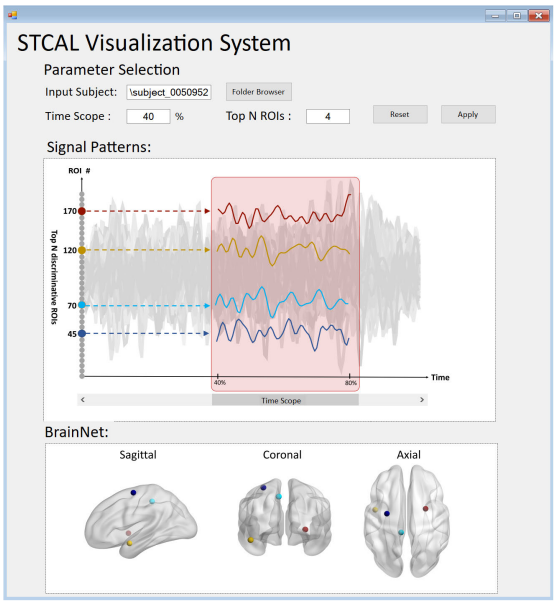
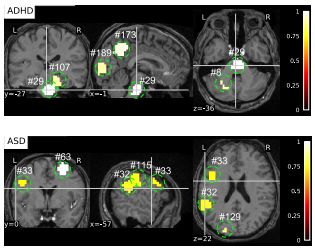


Fig. 15. Visualization demo of the STCAL visualization system.



Five top predicted ADHD-specific ROIs

ID	Region	PMID	Score
29	Right Cingulate Gyrus	20008822	0.989
173	Left Superior Frontal Gyrus	31265198	0.976
8	Left Postcentral Gyrus	23800705	0.914
189	Left Fusiform Gyrus	25938917	0.909
107	Right Middle Temporal Gyrus	16129560	0.891

Five top predicted ASD-specific ROIs

ID	Region	PMID	Score
63	Left Inferior Temporal Gyrus	30425664	0.998
129	Left Middle Temporal Gyrus	31729198	0.883
32	Right Temporal Pole	20858472	0.827
115	Right Precentral Gyrus	32954337	0.806
33	Left Superior Temporal Gyrus	16806505	0.757

Fig. 16. Showcase for clinical interpretation analysis in terms of patients' group.

from the same group. As shown in Fig. 16, we showcase the five top disorder-specific ROIs, which are significant in the case of the group while unremarkable in the control group. Through manual investigation, all five top predicted disease-specific ROIs have been validated to have associations with the neurological manifestations of ADHD and ASD by previous literature. For example, a significant volumetric decrease in the right cingulate gyrus (29) was found among the treated group with ADHD [59]. The subjects with ASD were reported to have increased gray matter volume in the left inferior temporal gyrus in subjects (63) [60]. This result demonstrates that the data-driven outcomes of STCAL are compatible with the previous clinical findings.

V. Conclusion

In this article, we have proposed an STCAL framework to identify multiple mental disorders from spatial–temporal fMRI data via the co-attention learning mechanism. The SCA module is proposed to capture salient local patterns by aggregating the local attention representations from sliding window clusters. Based on the co-attention mechanism, we introduce a novel GCA module to learn the fine-grained co-attention representations with the guidance of spatial–temporal attention matching. The SCA and GCA modules are proposed to capture salient local patterns and fine-grained co-attention representation, respectively. The comprehensive experiments on three real-world datasets have validated the effectiveness and reliability of STCAL. The following particular advantages have also been highlighted by the results of the experiment.

1) STCAL is composed mainly of the proposed SCA and GCA modules, which provide us with important properties such as extensibility, decoupling, and robustness so as to further explore, refine, and reconstruct the promising network structures according to the target goal (see Section IV-D).
2) Based on the spatial–temporal attention scores learned by STCAL, it is demonstrated that STCAL can tolerate at most 50% of feature-based pruning for key ROIs and time frames while maintaining stable accuracy and strong robustness (see Section IV-F).
3) Thanks to the proposed co-attention mechanism, the powerful interpretability allows STCAL to capture the salient local dynamic patterns by automatically locating the discriminative ROIs with respect to a given time scope (see Section IV-G).

To the best of our knowledge, this is the first attempt to develop co-attention learning framework for time-series fMRI data for diagnosing mental disorders. We expect that the current research can contribute to the development of an effective CAD system for achieving personalized diagnosis and clinical decision-making. This work aims to address the diagnosis of mental disorders from resting-state fMRI data. Since co-attention learning is adopted in multimodal tasks, we believe that it also holds great promise in the extension to other MRI tasks. For example, spatial-level features can be replaced with structural MRI features for multimodality fusion on both functional and structural MRI data.

References

[1] A. Thapar and M. Rutter, "Neurodevelopmental disorders," in *Rutter's Child and Adolescent Psychiatry*. Amsterdam, The Netherlands: Elsevier, 2015, pp. 31–40.
[2] C. Kieling et al., "Child and adolescent mental health worldwide: Evidence for action," *Lancet*, vol. 378, no. 9801, pp. 1515–1525, 2011.
[3] T. Chung, J. Cornelius, D. Clark, and C. Martin, "Greater prevalence of proposed ICD-11 alcohol and cannabis dependence compared to ICD-10, DSM-IV, and DSM-5 in treated adolescents," *Alcoholism: Clin. Experim. Res.*, vol. 41, no. 9, pp. 1584–1592, Sep. 2017.
[4] F. V. Farahani, W. Karwowski, and N. R. Lighthall, "Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review," *Frontiers Neurosci.*, vol. 13, p. 585, Jun. 2019.
[5] Z.-A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan, "Identifying autism spectrum disorder from resting-state fMRI using deep belief network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2847–2861, Jul. 2020.
[6] C. Dou, S. Zhang, H. Wang, L. Sun, Y. Huang, and W. Yue, "ADHD fMRI short-time analysis method for edge computing based on multi-instance learning," *J. Syst. Archit.*, vol. 111, Dec. 2020, Art. no. 101834.
[7] J. Liu, J. Ji, G. Xun, and A. Zhang, "Inferring effective connectivity networks from fMRI time series with a temporal entropy-score," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5993–6006, Oct. 2022.
[8] L. Xia et al., "Mulhita: A novel multiclass classification framework with multibranch lstm and hierarchical temporal attention for early detection of mental stress," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 6, 2022, doi: 10.1109/TNNLS.2022.3159573.
[9] W. Gao, H. Zhu, K. Giovanello, and W. Lin, "Multivariate network-level approach to detect interactions between large-scale functional systems," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Midtown Manhattan, NY, USA: Springer, 2010, pp. 298–305.

[10] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[11] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, Oct. 2022.

[12] M. Greicius, "Resting-state functional connectivity in neuropsychiatric disorders," *Current Opinion Neurol.*, vol. 21, no. 4, pp. 424–430, Aug. 2008.

[13] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6281–6290.

[14] H. Zhang et al., "Multiscale visual-attribute co-attention for zero-shot image recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 17, 2021, doi: 10.1109/TNNLS.2021.3132366.

[15] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[16] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[17] J. C. Daza and A. Rueda, "Classification of Alzheimer's disease in MRI using visual saliency information," in *Proc. IEEE 11th Colombian Comput. Conf. (CCC)*, Sep. 2016, pp. 1–7.

[18] O. Ben-Ahmed, F. Lecellier, M. Paccalin, and C. Fernandez-Maloigne, "Multi-view visual saliency-based MRI classification for alzheimer's disease diagnosis," in *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2017, pp. 1–6.

[19] Y. Liu, L. Xu, J. Li, J. Yu, and X. Yu, "Attentional connectivity-based prediction of autism using heterogeneous rs-fMRI data from CC200 atlas," *Experim. Neurobiol.*, vol. 29, no. 1, p. 27, 2020.

[20] X. Dai, X. Kong, X. Liu, J. B. Lee, and C. Moore, "Dual-attention recurrent networks for affine registration of neuroimaging data," in *Proc. SIAM Int. Conf. Data Mining*, 2020, pp. 379–387.

[21] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, Oct. 2021.

[22] K. Niu et al., "Multichannel deep attention neural networks for the classification of autism spectrum disorder using neuroimaging and personal characteristic data," *Complexity*, vol. 2020, pp. 1–9, Jan. 2020.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[24] J. Zhang, B. Zheng, A. Gao, X. Feng, D. Liang, and X. Long, "A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification," *Magn. Reson. Imag.*, vol. 78, pp. 119–126, May 2021.

[25] Y. Chen et al., "Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 14, 2022, doi: 10.1109/TNNLS.2022.3154755.

[26] Y. Zhang et al., "3D multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification," *IEEE Trans. Med. Imag.*, vol. 40, no. 6, pp. 1618–1631, Jun. 2021.

[27] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 14, 2022, doi: 10.1109/TNNLS.2022.3219551.

[28] C. Lian, M. Liu, L. Wang, and D. Shen, "Multi-task weakly-supervised attention network for dementia status estimation with structural MRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 4056–4068, Aug. 2022.

[29] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: Classification of morphological patterns using adaptive regional elements," *IEEE Trans. Med. Imag.*, vol. 26, no. 1, pp. 93–105, Jan. 2006.

[30] C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans, "Enhancement of MR images using registration for signal averaging," *J. Comput. Assist. Tomogr.*, vol. 22, no. 2, pp. 324–333, 1998.

[31] J. Zhang, Y. Gao, Y. Gao, B. C. Munsell, and D. Shen, "Detecting anatomical landmarks for fast Alzheimer's disease diagnosis," *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2524–2533, Dec. 2016.

[32] D. S. Bassett, B. G. Nelson, B. A. Mueller, J. Camchong, and K. O. Lim, "Altered resting state complexity in schizophrenia," *NeuroImage*, vol. 59, no. 3, pp. 2196–2207, Feb. 2012.

[33] L. Wang et al., "Cognitive and behavior deficits in sickle cell mice are associated with profound neuropathologic changes in hippocampus and cerebellum," *Neurobiol. Disease*, vol. 85, pp. 60–72, Jan. 2016.

[34] H. Shen, L. Wang, Y. Liu, and D. Hu, "Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI," *NeuroImage*, vol. 49, no. 4, pp. 3110–3121, Feb. 2010.

[35] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, and D. Rueckert, "Multiple instance learning for classification of dementia in brain MRI," *Med. Image Anal.*, vol. 18, no. 5, pp. 808–818, 2014.

[36] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 880–893, Apr. 2018.

[37] Z. Mao et al., "Spatio–temporal deep learning method for ADHD fMRI classification," *Inf. Sci.*, vol. 499, pp. 1–11, Oct. 2019.

[38] D. Pena et al., "Quantifying neurodegenerative progression with DeepSymNet, an end-to-end data-driven approach," *Frontiers Neurosci.*, vol. 13, p. 1053, Oct. 2019.

[39] C. Lian, M. Liu, Y. Pan, and D. Shen, "Attention-guided hybrid network for dementia diagnosis with structural MR images," *IEEE Trans. Cybern.*, vol. 52, no. 4, pp. 1992–2003, Apr. 2020.

[40] Z. Wang et al., "Distribution-guided network thresholding for functional connectivity analysis in fMRI-based brain disorder identification," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 4, pp. 1602–1613, Apr. 2021.

[41] J. Wang, Q. Wang, H. Zhang, J. Chen, S. Wang, and D. Shen, "Sparse multiview task-centralized ensemble learning for ASD diagnosis based on age- and sex-related functional connectivity patterns," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 3141–3154, Aug. 2019.

[42] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3464–3473.

[43] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 10012–10022.

[44] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, pp. 150–157, Jul. 2001.

[45] J. Gu et al., "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[46] G. S. Sidhu, N. Asgarian, R. Greiner, and M. R. G. Brown, "Kernel principal component analysis for dimensionality reduction in fMRI-based diagnosis of ADHD," *Frontiers Syst. Neurosci.*, vol. 6, p. 74, Nov. 2012.

[47] F. Almuqhim and F. Saeed, "ASD-SAENet: A sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data," *Frontiers Comput. Neurosci.*, vol. 15, p. 27, Apr. 2021.

[48] Z. Sherkatghanad et al., "Automated detection of autism spectrum disorder using a convolutional neural network," *Frontiers Neurosci.*, vol. 13, p. 1325, Jan. 2020.

[49] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, and F. Saeed, "ASD-DiagNet: A hybrid learning approach for detection of autism spectrum disorder using fMRI data," *Frontiers Neuroinform.*, vol. 13, p. 70, Nov. 2019.

[50] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage, Clin.*, vol. 17, pp. 16–23, Jan. 2018.

[51] J. Liu and J. Ji, "Classification method of fMRI data based on broad learning system," *J. ZheJiang Univ. Eng. Sci.*, vol. 55, no. 7, pp. 1270–1278, 2021.

[52] T. Chen et al., "The development of a practical artificial intelligence tool for diagnosing and evaluating autism spectrum disorder: Multicenter study," *JMIR Med. Informat.*, vol. 8, no. 5, p. e15767, 2020.

[53] M. A. Aghdam, A. Sharifi, and M. M. Pedram, "Diagnosis of autism spectrum disorders in young children based on resting-state functional magnetic resonance imaging data using convolutional neural networks," *J. Digit. Imag.*, vol. 32, no. 6, pp. 899–918, Dec. 2019.

[54] Y. Zhao, H. Dai, W. Zhang, F. Ge, and T. Liu, "Two-stage spatial temporal deep learning framework for functional brain network modeling," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1576–1580.

[55] L. Zou, J. Zheng, C. Miao, M. J. Mckeown, and Z. J. Wang, "3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI," *IEEE Access*, vol. 5, pp. 23626–23636, 2017.
[56] A. A. Pulini, W. T. Kerr, S. K. Loo, and A. Lenartowicz, "Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis," *Biol. Psychiatry, Cognit. Neurosci. Neuroimaging*, vol. 4, no. 2, pp. 108–120, Feb. 2019.
[57] M. R. Brown et al., "ADHD-200 global competition: Diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements," *Frontiers Syst. Neurosci.*, vol. 6, p. 69, Sep. 2012.
[58] Y. Li, J. Liu, Z. Tang, and B. Lei, "Deep spatial–temporal feature fusion from adaptive dynamic functional connectivity for MCI identification," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2818–2830, Sep. 2020.
[59] N. Makris et al., "Anterior cingulate volumetric alterations in treatment-naive adults with ADHD: A pilot study," *J. Attention Disorders*, vol. 13, no. 4, pp. 407–413, Jan. 2010.
[60] J. Cai, X. Hu, K. Guo, P. Yang, M. Situ, and Y. Huang, "Increased left inferior temporal gyrus was found in both low function autism and high function autism," *Frontiers Psychiatry*, vol. 9, p. 542, Oct. 2018.

Rui Liu received the B.S. degree in intelligence science and technology from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong.

His current research interests include machine learning, multitask/modality learning, and medical images diagnosis and applied deep learning.

Zhi-An Huang received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2021.

He is currently a Research Fellow with the Research Office, City University of Hong Kong (Dongguan), Hong Kong. His research interests include artificial intelligence, machine learning, bioinformatics, and medical imaging analysis.

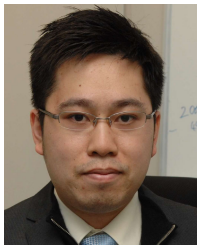
Yao Hu received the B.S. degree in mining engineering and the M.Sc. degree in control science and engineering from the China University of Mining and Technology, Xuzhou, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong.

His current research interests include machine learning, transfer learning, and federated learning.

Zexuan Zhu (Senior Member, IEEE) received the B.S. degree in computer science and technology from Fudan University, China, in 2003, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2008.

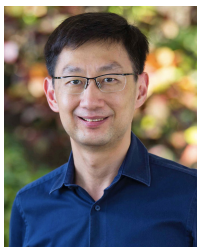
He is currently a Professor with the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China. His research interests include computational intelligence, machine learning, and bioinformatics.

Dr. Zhu was an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. He is currently an Associate Editor of the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE and the Chair of the IEEE CIS Emergent Technologies Task Force on Memetic Computing.



Ka-Chun Wong received the B.Eng. degree in computer engineering and the M.Phil. degree from the Chinese University of Hong Kong, Hong Kong, in 2008 and 2010, respectively, and the Ph.D. degree from the Department of Computer Science, University of Toronto, Toronto, ON, Canada, in 2015.

He was an Associate Professor with the City University of Hong Kong, Hong Kong. His current research interests include bioinformatics, computational biology, evolutionary computation, data mining, machine learning, and interdisciplinary research.



Kay Chen Tan (Fellow, IEEE) received the B.Eng. degree (Hons.) and the Ph.D. degree from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively.

He is currently a Chair Professor with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong.

Dr. Tan was the Editor-in-Chief of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, and currently serves on the Editorial Board member of 10+ journals. He is currently the Vice-President (Publications) of the IEEE Computational Intelligence Society, an Honorary Professor at the University of Nottingham, Nottingham, U.K., and the Chief Coeditor of Springer Book Series on Machine Learning: Foundations, Methodologies, and Applications.