

Diffusion Kernel Attention Network for Brain Disorder Classification

Jianjia Zhang^{ID}, Luping Zhou^{ID}, Senior Member, IEEE, Lei Wang^{ID}, Mengting Liu,
and Dinggang Shen^{ID}, Fellow, IEEE

Abstract—Constructing and analyzing functional brain networks (FBN) has become a promising approach to brain disorder classification. However, the conventional successive construct-and-analyze process would limit the performance due to the lack of interactions and adaptivity among the subtasks in the process. Recently, Transformer has demonstrated remarkable performance in various tasks, attributing to its effective attention mechanism in modeling complex feature relationships. In this paper, for the first time, we develop Transformer for integrated FBN modeling, analysis and brain disorder classification with rs-fMRI data by proposing a Diffusion Kernel Attention Network to address the specific challenges. Specifically, directly applying Transformer does not necessarily admit optimal performance in this task due to its extensive parameters in the attention module against the limited training samples usually available. Looking into this issue, we propose to use kernel attention to replace the original dot-product attention module in Transformer. This significantly reduces the number of parameters to train and thus alleviates the issue of small sample while introducing a non-linear attention mechanism to model complex functional connections. Another limit of Transformer for FBN applications is that it only considers pair-wise interactions between directly connected brain regions but ignores the important indirect connections. Therefore, we further explore diffusion process over the kernel attention to incorporate wider interactions among indirectly connected brain regions. Extensive experimental study is conducted on ADHD-200 data set for ADHD classification and on ADNI data set for Alzheimer's disease classification, and the results demonstrate the superior performance of the proposed method over the competing methods.

Index Terms—Attention network, brain disease classification, brain network, kernel, diffusion process, transformer.

I. INTRODUCTION

FUNCTIONAL brain network (FBN) construction and analysis with resting-state functional magnetic resonance imaging (rs-fMRI) hold great promise for brain disease classification [1]–[6]. Rs-fMRI, as an in vivo brain functional imaging technique, measures the oscillations of blood-oxygen-level-dependent (BOLD) signals which reflect neuronal activities when the subject being scanned is in natural rest [7], [8]. By splitting the brain into separate regions of interest (ROIs) [9], [10] as nodes and evaluating the connectivity between these nodes with the associated rs-fMRI time series, FBN can be constructed for further analysis. Such FBN reflects the inherent functional interactions among different brain regions and may present characteristic abnormal patterns in the subjects affected by brain disorders, e.g., Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD) and epilepsy. These disease-specific insights on the connectivity patterns of FBN could serve as features for early brain disorder classification and also help to study its pathogenesis [5], [11], [12], and the former is the focus of this study.

Traditional FBN-based methods for brain disorder classification usually conduct FBN construction and analysis in two separate steps. In the first step, typical FBNs are built by identifying brain nodes and inferring the connectivities between them [13]. Specifically, FBN nodes are often defined as anatomically separated brain regions [14] or alternatively as latent components obtained by data-driven methods, e.g., independent component analysis and clustering-based methods [13]. With the identified nodes, the strength of functional connectivity between a pair of nodes is conventionally measured by the co-varying pattern of the averaged time series associated with the two nodes, such as the correlation [5], [13], partial correlation [15] or representation coefficients [16]. In the second step, various features, including hand-crafted graph features [16], [17], principal component coefficients [18], [19], or recent learning-based deep features [8], [20], [21], can be extracted from FBNs as input for classification. Extensive studies following this approach have demonstrated promising results in brain disorder understanding and classification.

However, we argue that the traditional two-step approach may not be optimal considering the following two issues. First,

Manuscript received 5 March 2022; revised 1 April 2022, 7 April 2022, and 18 April 2022; accepted 23 April 2022. Date of publication 26 April 2022; date of current version 30 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62131015 and Grant 62101611, in part by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 21010502600, and in part by the Natural Science Foundation of Guangdong Province under Grant 2022A1515011375. (Corresponding author: Dinggang Shen.)

Jianjia Zhang and Mengting Liu are with the School of Biomedical Engineering, Sun Yat-sen University, Shenzhen 518107, China (e-mail: zhangjj225@mail.sysu.edu.cn; liumt55@mail.sysu.edu.cn).

Luping Zhou is with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: luping.zhou@sydney.edu.au).

Lei Wang is with the School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail: leiw@uow.edu.au).

Dinggang Shen is with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China, and also with Shanghai United Imaging Intelligence Company Ltd., Shanghai 200230, China (e-mail: dinggang.shen@gmail.com).

Digital Object Identifier 10.1109/TMI.2022.3170701

its performance is probably not optimal due to the lack of communication between the two steps. Specifically, the FBNs constructed in the first step do not necessarily work well with the feature extraction methods in the second step since there are no interactions between them and there is no unified objective [22]. Although many recent works [8], [20], [21] shift to extracting deep learning-based features from FBNs, the information lost in the previous FBN construction step can in no way be recovered. Second, the traditional way of FBN construction and analysis would limit the applicability. Usually rich experience and domain knowledge in both machine learning and neurology are required to select an appropriate combination of FBN construction and analysis methods. Such experience and knowledge may not be always accessible considering that it requires the joint efforts of machine learning experts and neurological physicians and can hardly generalize to different brain disorders.

To break the limitations of the traditional two-step approach, it is desirable to develop an integrated framework for the joint optimization of FBN construction and analysis. A handful of studies have exploited this pioneering field. Several works [23]–[26] directly learn adaptive feature representations from rs-fMRI data with deep architectures. Specifically, [23] and [24] discovered the spatio-temporal patterns in rs-fMRI for AD and ADHD classification, respectively, with auto-encoder architectures. In contrast, restricted Boltzmann machine was exploited in [25], [26] to interpret fMRI time courses explicitly and explore latent features. However, these works focused on feature extraction from rs-fMRI but did not integrate FBN modeling in the deep architectures. The recent works in [21], [27] introduced certain flexibility into the FBN construction process within graph convolution networks (GCN) framework. Specifically, a binary FBN mask is first predefined by using diffusion tensor imaging (DTI) data, and only the functional connection strength between brain ROIs is optimized in [21] while [27] initializes the topology of the graph with structural DTI network and iteratively updates it by incorporating the functional influences from rs-fMRI data. In contrast, our work utilizes sole rs-fMRI data and proposes to construct fully adaptive FBN within the Transformer framework. GCN was also adopted in [28] to extract both spatial and temporal dynamics in rs-fMRI data. In that work, FBN construction and analysis were simultaneously conducted. However, the model was intentionally designed for gender and age prediction, which is different from the brain disorder classification task focused in our work. Moreover, as will be demonstrated later in the experiments, our proposed method outperforms [28] when both are applied to brain disorder classification tasks.

The advent of Transformer [29] sheds new lights on designing a unified deep architecture for FBN construction and analysis. Attributing to its exceptional capability to model long-range dependencies in sequential data by the self-attention mechanism, Transformer has demonstrated impressive performance in various vision tasks. Vision Transformer is firstly proposed in [30] by applying a pure transformer directly to image patches for image recognition. Since then, a variety of Transformer variants have been developed and successfully applied in many vision tasks, e.g., image

classification [31], [32], object detection [33], [34] and image generation [35]. Refer to [36], [37] for comprehensive surveys on Transformers in computer vision. Reviewing these literature, our study finds that the key concepts and components of Transformer are well aligned with the requirements of the FBN construction and analysis task. Specifically, rs-fMRI time series is typical sequential data which well fits Transformer framework, and the embedding functions and self-attention module in Transformer could extract discriminative features from the time series of brain ROIs and measure their pair-wise similarities as connection strength. In this case, the attention matrix over all pairs of brain ROIs results in a FBN representation. Moreover, the multi-head mechanism in Transformer could construct multi-view FBNs to extract more comprehensive information for further analysis. The stacked attention modules in Transformer encoder can be interpreted either as hierarchical FBNs or multi-level feature extraction layers on the bottom FBN construction layer. The final output features learned by these modules are fed into a classification layer for classification. In this case, the whole architecture, including the FBN construction, analysis and classification layers, are inherently integrated. During the training process, all the modules in the architecture can be optimized in an end-to-end manner, and they communicate and negotiate with each other to reach an overall unified objective, finally leading to an improved classification accuracy.

Nevertheless, a direct application of the original Transformer to FBN based brain disorder classification may still not be optimal considering the following issues:

- The architecture of Transformer, especially the parameterized feature projections in the attention module, consists of parameters whose number increases quadratically with respect to the input dimensions [35]. Therefore, a sufficiently large number of training samples are required to train the architecture. However, collecting such large data sets in brain disorder classification tasks is exceptionally time-consuming and labour-expensive, if not impossible. In this case, it is desirable to reduce the number of parameters while maintaining or even improving the effectiveness of Transformer;
- The non-linear relationships between the time series of brain ROIs will not be fully exploited if the original Transformer is directly applied. The dot-product attention mechanism in the original Transformer can only model the linear correlations between the projected time series of brain ROIs [38]. However, the interactions between these brain regions are known to be complex and not necessarily only linear [39]. Modeling additional higher order relationships between the time series of brain regions could better reveal their underlying functional communications and in turn provide more clues for brain disorder classification;
- The attention mechanism of Transformer only models pairwise relationships between brain ROIs, ignoring wider range of interactions among indirectly connected brain ROIs. Recent studies have demonstrated that one brain region may not only interact with its directly connected neighbors, but also implicitly communicate

with the regions connected to its neighbors [15], [21]. Therefore, considering wider range of interactions could gain more comprehensive insights on the connections among either directly or indirectly connected brain ROIs for brain disorder classification.

To address the issues above, this paper proposes a diffusion kernel attention network for brain disorder classification. With the encoder network of Transformer as the backbone, the proposed network incorporates two major modifications by considering the specific properties of the task in our work. On the one hand, in order to address the first two issues mentioned above, i.e., huge number of parameters and incapability of modeling non-linear interactions, a kernel attention mechanism is proposed to replace the original attention module. Specifically, as previously mentioned, the dot-product attention in the original transformer essentially calculates the linear dependencies between features as similarity measures. From this perspective, we propose to use a kernel function to replace the original dot-product operation in the attention module, leading to a kernel attention mechanism. As a classical method in many machine learning techniques, kernel functions enable feature similarity measures by implicitly mapping data onto a high-dimensional space and then computing the inner products between pairs of data in that space. With the kernel trick, the computation of the self-attention requires much fewer parameters, which can be automatically learned during model training. This not only significantly reduces the number of parameters to optimize, alleviating the issue of small sample, but also introduces non-linear attention mechanism to model complex interactions between brain regions. In addition, it admits high flexibility in modeling various nonlinearities by choosing different kernels. Although interpreting and improving the attention module from the kernel perspective was attempted in some recent works [35], [38], [40], those works focus on reformulating the attention module with the soft-max to derive fast computations in natural language processing [38], image recognition [35] or segmentation [40]. In contrast, our work, to the best of our knowledge, is the first work adapting the attention module with kernels from the perspective of reducing the number of parameters for FBN modeling and analysis.

On the other hand, in order to address the third issue and incorporate wider interactions among indirectly connected brain regions, diffusion process is further incorporated into our work on top of the kernel attention. The functionality of brain nodes does not only lie in their communication with their direct neighbors, but also their indirect network-level connectomes [41], [42]. Therefore, besides inferring the pairwise connections between brain nodes with the proposed kernel attention, modeling wider functional interactions could additionally reveal the brain's functional properties. To this end, random walks, as a commonly used Diffusion process method [21], [43], [44], is integrated into our work to model wider node interactions in brain networks. With the help of random walks, the information of one brain region could propagate on the network from itself to faraway neighbors through the indirect connected paths. During this process, the feature representation of one node could be refined by incorporating the features of its own and its indirectly connected neighbors',

presenting richer connection characteristics for brain disorder classification.

The major contributions of this paper can be summarized as follows. First, as far as we are aware, this is the first work adapting Transformer for integrated FBN construction and analysis, resulting in a unified end-to-end framework for brain disorder classification. This makes the approach less experience-dependent, but more flexible and applicable. Second, after insightful analysis of the architecture of the original Transformer and specific characteristics of the brain disorder classification task in this paper, a kernelized attention mechanism is proposed. With this, the improved Transformer is able to model non-linear relationships between brain regions while reducing the number of samples required to train, alleviating the adverse effects of small sample problem. Furthermore, diffusion process is performed on the top of the kernel attention to incorporate wider range of interactions between indirectly connected brain regions. Lastly, the proposed method is evaluated in ADHD and AD classification tasks, demonstrating its superior performance over the competing methods.

II. METHOD

A. Overall Architecture

As illustrated in Fig. 1(a), the overall architecture of the proposed model consists of a frontal linear input projection module, a position encoding module, L stacked Transformer encoder layers, an average pooling layer and a fully connected layer with softmax at the end as the classifier. Specifically, let $X_{raw} \in \mathbb{R}^{N \times T \times C}$ denote an input raw rs-fMRI time series of N ROIs, T frames and C feature channels, corresponding to N tokens of $T \times C$ dimensions. X_{raw} is first embedded by a linear projection matrix $W_p \in \mathbb{R}^{C \times D}$, i.e., $X_p = X_{raw} W_p$, and $X_p \in \mathbb{R}^{N \times T \times D}$ to fit a D -dimensional encoder layer. Note that, different from the vanilla Transformer where each token is embedded into a D -dimensional vector, in our model, each token is still represented by a $T \times D$ matrix to keep the temporal dimension for further analysis and save projection parameters. This additional T dimension will be eliminated by the global average pooling layer following the stacked encoder layers, as introduced later. Following the linear projection module, the position encoding module incorporates ROI indexes into the projected rs-fMRI signals as spatial identity by following [29]. Specifically, the position encoding module maps each ROI index to a real valued D -dimensional vector, i.e., $\mathbb{N} \rightarrow \mathbb{R}^D$, and the resulting encoding matrix, denoted as $E \in \mathbb{R}^{N \times T \times D}$, is defined as

$$\begin{aligned} E(pos, :, 2dim) &= \sin(pos/10000^{2dim/D}) \\ E(pos, :, 2dim+1) &= \cos(pos/10000^{2dim/D}), \end{aligned} \quad (1)$$

where $pos \in [0, \dots, N-1]$ is the brain ROI index, the symbol ':' refers to all indexes in the corresponding dimension, and $dim \in [0, \dots, \frac{D}{2}-1]$ indicates the index of the feature dimension. Then E is added to the projected input X_p , i.e., $X_0 = X_p + E$. The resulting X_0 is fed into L stacked encoder layers. Each Transformer encoder layer is essentially a feature

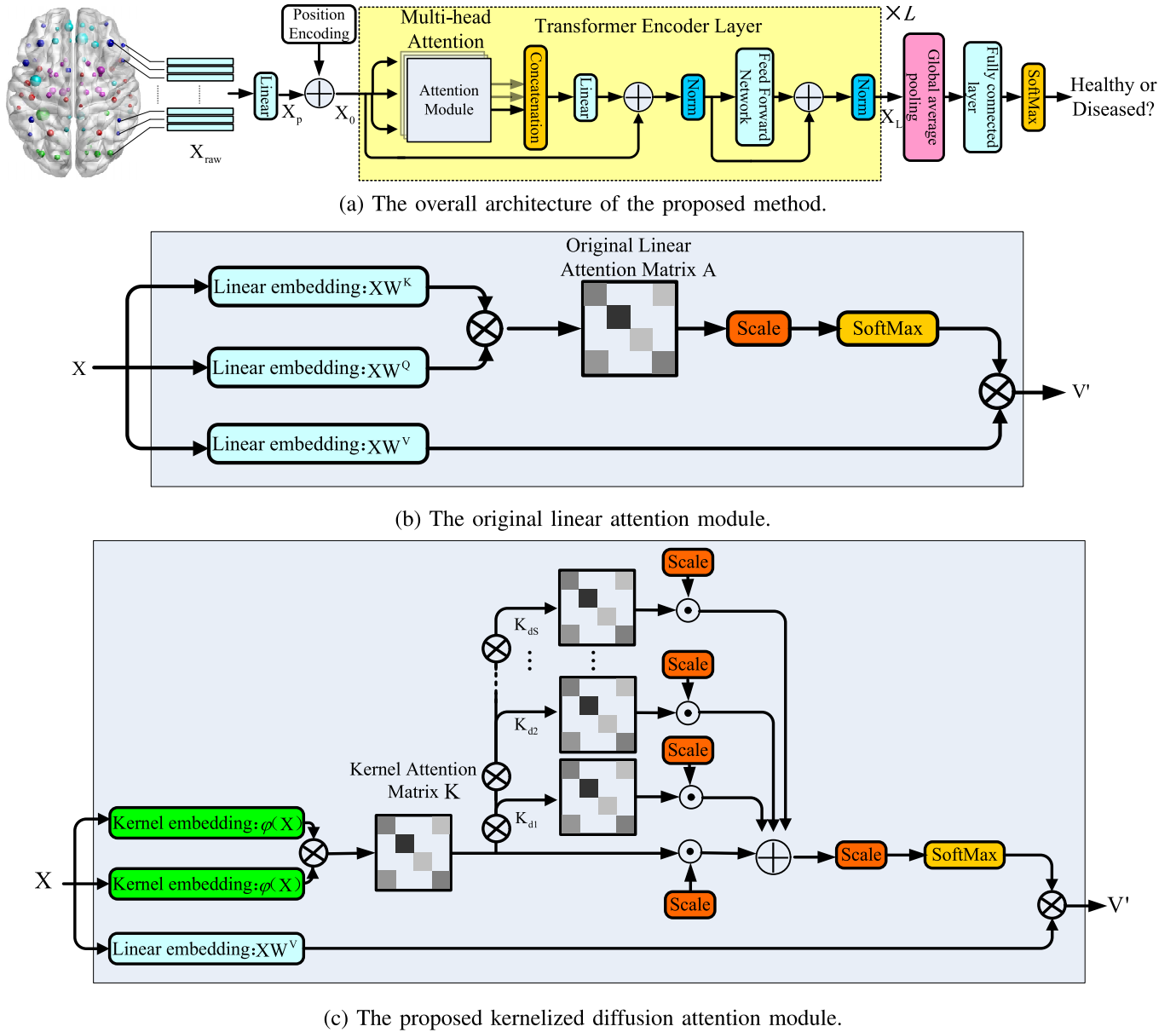


Fig. 1. (a) The overall architecture of the proposed method; (b) the illustration of the original linear attention module; and (c) the proposed diffusion kernel attention.

transformation function, which can be formulated as

$$X_l = f_l(a_l(X_{l-1}), X_{l-1}), \quad l = 1, \dots, L. \quad (2)$$

Specifically, as illustrated in Fig. 1 (a), f_l transforms the input feature representations of ROIs, X_{l-1} , by applying a multi-head self-attention function, denoted as $a_l(\cdot)$, a linear projection layer and a fully connected feed-forward network. Residual connections and normalizations are also employed around the multi-head self-attention and the fully connected feed-forward network modules as in [29]. The multi-head attention function $a_l(\cdot)$ consists of multiple parallel self-attention modules with the same input, and their outputs are concatenated into a joint feature for further processing. Investigating the structure of the self-attention module and adapting it to fit brain disorder classification is the focus of this paper and will be elaborated in detail in the following

sections. The output of the last stacked transformer encoder layer, i.e., $X_L \in \mathbb{R}^{N \times T \times D}$, is condensed along the first and the second dimensions by global average pooling, resulting in a D -dimensional vector to be fed into the fully connected and softmax layers for classification scores. Note that the dimension D of Transformer layer can be either kept fixed as in the original Transformer [29] or set differently across the L layers as in this paper to save parameters while introducing flexibility.

B. Revisiting Self-Attention Module

The multi-head self-attention module is the key block of Transformer. As illustrated in Fig. 1(b), a self-attention module in the l -th layer takes $X_{l-1} \in \mathbb{R}^{N \times T \times D}$ as the input and generates the three key components of self-attention, i.e., the

queries $\mathbf{Q} \in \mathbb{R}^{N \times T \times D_k}$, the keys $\mathbf{K} \in \mathbb{R}^{N \times T \times D_k}$ and the values $\mathbf{V} \in \mathbb{R}^{N \times T \times D_v}$ by linear projections of \mathbf{X}_{l-1} with three trainable projection matrices $\mathbf{W}^Q \in \mathbb{R}^{D \times D_k}$, $\mathbf{W}^K \in \mathbb{R}^{D \times D_k}$, $\mathbf{W}^V \in \mathbb{R}^{D \times D_v}$, respectively, as follows:

$$\begin{aligned}\mathbf{Q} &= \mathbf{X}_{l-1} \mathbf{W}^Q \\ \mathbf{K} &= \mathbf{X}_{l-1} \mathbf{W}^K \\ \mathbf{V} &= \mathbf{X}_{l-1} \mathbf{W}^V.\end{aligned}\quad (3)$$

D_k and D_v are set as D/H , where H is the number of heads which will be introduced later. In order to facilitate the following attention calculation, both \mathbf{Q} and \mathbf{K} are reshaped into $N \times T D_k$ dimensional matrices. With these three components, the self-attention in the l -th layer is defined as:

$$a_l(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V}' = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{T D_k}}\right)\mathbf{V}, \quad (4)$$

where $\text{softmax}(\cdot)$ refers to a row-wise softmax function. Let us denote the i -th rows of \mathbf{V}' , \mathbf{Q} , \mathbf{K} , and \mathbf{V} as \mathbf{v}'_i , \mathbf{q}_i , \mathbf{k}_i , and \mathbf{v}_i , respectively. We have:

$$\mathbf{v}'_i = \frac{1}{\sum_{j=1}^N \exp\left(\frac{\langle \mathbf{q}_i, \mathbf{k}_j \rangle}{\sqrt{T D_k}}\right)} \sum_{j=1}^N \exp\left(\frac{\langle \mathbf{q}_i, \mathbf{k}_j \rangle}{\sqrt{T D_k}}\right) \mathbf{v}_j, \quad (5)$$

where $\frac{1}{\sum_{j=1}^N \exp\left(\frac{\langle \mathbf{q}_i, \mathbf{k}_j \rangle}{\sqrt{T D_k}}\right)}$ is just a total scalar for normalization.

Eq. (5) indicates that the self-attention mechanism depends on the dot-product, which is essentially a similarity measure, between pairwise ROI features \mathbf{q}_i and \mathbf{k}_j [38]. The corresponding matrix $\mathbf{Q}\mathbf{K}^T$, denoted as \mathbf{A} , is a $N \times N$ dimensional matrix. Referring to Eq. (3), $A_{i,j}$ can be rewritten as follows:

$$\begin{aligned}A_{i,j} &= \langle \mathbf{q}_i, \mathbf{k}_j \rangle \\ &= \langle \mathbf{x}_i \mathbf{W}^Q, \mathbf{x}_j \mathbf{W}^K \rangle\end{aligned}\quad (6)$$

where \mathbf{x}_i and \mathbf{x}_j denote the i -th and j -th rows of \mathbf{X}_{l-1} . Reviewing Eq. (6), $A_{i,j}$ essentially measures the dependency between the projected features of the i -th and the j -th ROIs. Similar to the typical correlation method in traditional connection calculation methods, such dependency can also be considered as the connection strength between the pair of ROIs straightforwardly. In this case, the resulting matrix \mathbf{A} can be interpreted as a functional brain network.

The above section presents the operation of a single head and there are H heads operating in parallel. The outputs of the H heads are concatenated to form the output of the whole multi-head attention module, i.e., $a_l(\mathbf{X}_{l-1}) = [\mathbf{V}'_1; \mathbf{V}'_2; \dots; \mathbf{V}'_H] \in \mathbb{R}^{N \times T \times D}$.

C. Proposed Kernel Self-Attention Module

Although enjoying simplicity, the original self-attention module in Transformer has a large number of parameters. In particular, each of the projection matrices \mathbf{W}^Q and \mathbf{W}^K has as many as $D \times D_k$ parameters, which increases the number of training samples required for optimization. However, the training data are usually scarce in the area of medical image analysis, including brain disorder classification tasks. After a careful review of the operations in the original self-attention module, we identify that the attention $A_{i,j}$ in Eq. (6), i.e.,

$\langle \mathbf{x}_i \mathbf{W}^Q, \mathbf{x}_j \mathbf{W}^K \rangle$, is essentially a linear kernel between projected \mathbf{x}_i and \mathbf{x}_j . From this perspective, a non-linear kernel can be utilized as a substitute to measure their similarity as follows:

$$\begin{aligned}A_{i,j} &= \langle \mathbf{x}_i \mathbf{W}^Q, \mathbf{x}_j \mathbf{W}^K \rangle \\ &\Rightarrow \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \\ &= k(\mathbf{x}_i, \mathbf{x}_j) \\ &= \mathbf{K}_{i,j},\end{aligned}\quad (7)$$

where $\phi(\cdot)$ is an implicit kernel feature mapping function, which plays a similar role to the projection matrices \mathbf{W}^Q or \mathbf{W}^K . Such a kernel mapping implicitly maps the input feature \mathbf{x} onto a higher-dimensional feature space and enables complicated interaction modeling. With the help of this kernel trick, the inner product between the mapped features, i.e., $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$, is efficiently computed by a kernel function with very few parameters. For example, with a Radial Basis Function (RBF) kernel, $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ is obtained by simply calculating $\exp(-\beta \|\mathbf{x}_i - \mathbf{x}_j\|^2)$, where β is the kernel parameter. As illustrated in Fig. 1(c), the gram kernel matrix \mathbf{K} over all pairs of \mathbf{x}_i and \mathbf{x}_j acts as a kernel-based attention, which can be interpreted as a kernel brain network. An important advantage of adopting the proposed kernel attention is that modeling complicated interactions via a kernel function could better reveal the underlying connectome dynamics among brain regions for brain disorder classification. Moreover, with the proposed kernel attention, the parameters in the projection matrices \mathbf{W}^Q and \mathbf{W}^K can be avoided, which helps to reduce the scale of data required to train the model and thus improve the model robustness when training samples are scarce. In addition, the kernel parameter, i.e., the β in the RBF kernel mentioned above, is automatically learned during the model training process. As will be demonstrated in the experiment, the proposed method effectively improves the classification performance in brain disorder analysis, where training data is often limited and a robust model with fewer parameters is preferred. At the same time, switching \mathbf{W}^Q and \mathbf{W}^K to the kernel characterized by a single parameter β , to a certain extent, may impact the flexibility of modeling. However, it is noteworthy that using a nonlinear kernel instead of linear projections could somewhat mitigate this impact.

D. Diffusion Kernel Self-Attention Module

Another limitation of using the original self-attention module for brain network analysis lies in its incapability of modeling interactions between indirectly connected neighbors since it only infers direct pairwise connections between brain ROIs, as seen in Eq. (6). In this case, the interactions among multiple brain regions or between indirectly connected regions via intermediate regions may be neglected. However, recent studies have verified that these wider node interactions among indirectly connected brain regions are also critical in brain disease classification [41]–[43]. In order to incorporate such wider interactions into the proposed method, random walk process is further applied to the kernel attention module. Specifically, random walk process describes the phenomenon

and properties of random information flows along the brain network structure. This process is able to update the status of a brain node by considering the impacts of its direct and indirect neighbors. It is characterized by a transition matrix [44], which determines the probability of moving from one node to another. In the context of random walks on a brain network studied in this paper, the transition matrix, denoted as $\mathbf{P} \in \mathbb{R}^{N \times N}$, can be obtained by normalizing the kernel attention matrix \mathbf{K} obtained in the last section with summation of each row equal to one, i.e., $\mathbf{P}_{ij} = \frac{\mathbf{K}_{ij}}{\sum_{j=1}^N \mathbf{K}_{ij}}$. Intuitively, $\mathbf{P}_{i,j}$ is the probability of randomly moving from node i to node j within one step. S -step random walks refer to repetitively performing random walk process S times on the same brain network, which enables a node to reach its S -order neighbors and incorporate wider interactions. The kernel attention matrix \mathbf{K} is set as the initial brain connection state, and the state of the connection network after S -step random walks can be mathematically defined as:

$$\mathbf{K}_{dS} = \mathbf{P}^S \mathbf{K}, \quad (8)$$

where \mathbf{P}^S is the S -order power of \mathbf{P} . As illustrated in Fig. 1(c), in our work, various orders of diffusion kernel attentions are combined via a weighted sum, i.e., $\mathbf{K}_p = \sum_{j=0}^S \alpha_j \mathbf{P}^j \mathbf{K}$. The weights α_j , $j \in [0, 1, \dots, S]$, could be adaptively learned during training to better fit a specific task.

In sum, the proposed diffusion kernel self-attention module is able to adaptively extract non-linear dependencies among a wider range of brain ROIs. Such features could better reveal the properties of brain networks in terms of the complexity and range of interactions. As will be demonstrated in the experiments, the proposed diffusion kernel self-attention module could achieve better performance than the original linear attention mechanism in brain disorder classification.

III. EXPERIMENTS AND RESULTS

A. Materials and Preprocessing

Two rs-fMRI data sets, including ADHD-200 and ADNI data sets, are used to verify the effectiveness of the proposed kernelized diffusion attention network for brain disorder classification. The ADHD-200 data set is provided by the Neuro Bureau for differentiating attention deficit hyperactivity disorder (ADHD) from healthy control subjects. ADHD-200 consists of 768 training subjects and 197 test subjects¹ collected from eight independent imaging sites, including Kennedy Krieger Institute (KKI), NeuroIMAGE Sample (NeuroIMAGE), New York University Child Study Center (NYU), Oregon Health and Science University (OHSU), Peking University (PKU), University of Pittsburgh (UPittsburgh), Washington University (WashU) and Brown University. The summary of this data set is provided in Table II. The rs-fMRI data are processed with Athena pipeline. Specifically, the first four echo-planar imaging (EPI) volumes are removed for signal equilibrium and then slice timing, orientation and motion correction are performed. Each rs-fMRI image is co-registered

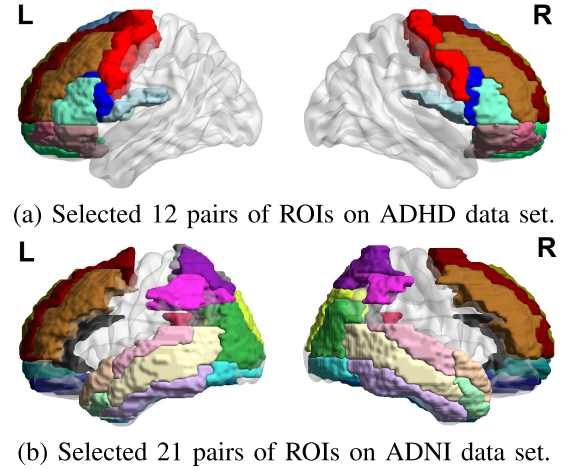


Fig. 2. Sagittal view of the selected ROIs in the left (denoted by 'L') and the right (denoted by 'R') hemispheres on (a) ADHD data set and (b) ADNI data set. Each pair of symmetrical regions shares a distinct color and the corresponding ROI names of different colors are provided in Table I.

to T1 image and warped into Montreal Neurological Institute (MNI) space at $4 \times 4 \times 4 \text{ mm}^3$ resolution. Detailed preprocessing descriptions and the processed time series are available at Neuro Bureau website.² The time series of 90 brain nodes in gray matter are extracted from the preprocessed data using the automated anatomical labeling (AAL) [14] atlas. Among them, 12 pairs of symmetrical regions of interest (ROIs) in the frontal lobe are selected in our study by following [45] since they have been confirmed to be highly correlated with ADHD in the literature [45]. These selected ROIs are illustrated in Fig. 2(a) with different colors, where each pair of symmetrical ROIs share the same color. The corresponding ROI names of different colors are provided in Table I.

The other data set is Alzheimer's Disease Neuroimaging Initiative (ADNI) data set downloaded from website³ with the aim of identifying mild cognitive impairment (MCI), which is early stage of Alzheimer's disease (AD), from healthy controls. Early MCI subjects are selected since they are more challenging and valuable for the early classification and treatment for AD patients [10], [46], [47]. There are 82 subjects in total with 44 early MCIs and 38 healthy controls. The data are acquired on a 3 Tesla (Philips) scanner with TR/TE set as 3000/30 ms and flip angle of 80° . Each series has 140 volumes, and each volume consists of 48 slices of 64×64 dimensional image matrices at $3.31 \times 3.31 \times 3.31 \text{ mm}^3$ resolution. The preprocessing is carried out using Statistical Parametric Mapping (SPM)-8⁴ and Data Processing Assistant for Resting-State fMRI Advanced-edition (DPARSFA) [48]. The first 10 volumes of each series are discarded for signal equilibrium. Slice timing, spatial smoothing, head motion correction, global drift removing and MNI space normalization are performed to ensure the consistency and reasonable scale of the rs-fMRI time series across different subjects. Participants with too







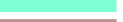

















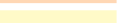
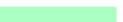



¹The labels of 26 subjects from Brown University in the test set are not released yet. Thus, they are not included in our performance evaluation.

²<http://neurobureau.projects.nitrc.org/ADHD200/Introduction.html>

³<http://adni.loni.usc.edu>

⁴<http://www.fil.ion.ucl.ac.uk/spm/software/>

TABLE I
THE FULL NAMES, ACRONYMS AND THE CORRESPONDING COLORS IN FIG. 2 OF THE SELECTED REGIONS OF INTEREST (ROIs) ON ADHD-200 AND ADNI DATA SETS

Full name of ROI	ROI Acronym	Is selected on ADHD-200	Is selected on ADNI	Color in Fig. 2
Precentral gyrus	PreCG	✓		
Superior frontal gyrus, dorsolateral	SFGdor	✓	✓	
Superior frontal gyrus, orbital part	ORBsup	✓		
Middle frontal gyrus	MFG	✓	✓	
Middle frontal gyrus, orbital part	ORBmid	✓		
Inferior frontal gyrus, opercular part	IFGoperc	✓		
Inferior frontal gyrus, triangular part	IFGtriang	✓		
Inferior frontal gyrus, orbital part	ORBinf	✓		
Rolandic operculum	ROL	✓		
Supplementary motor area	SMA	✓		
Superior frontal gyrus, medial	SFGmed	✓	✓	
Superior frontal gyrus, medial orbital	ORBsupmed	✓	✓	
Gyrus rectus	REC		✓	
Anterior cingulate and paracingulate gyri	ACG		✓	
Posterior cingulate gyrus	PCG		✓	
Hippocampus	HIP		✓	
Parahippocampal gyrus	PHG		✓	
Superior occipital gyrus	SOG		✓	
Middle occipital gyrus	MOG		✓	
Inferior occipital gyrus	IOG		✓	
Fusiform gyrus	FFG		✓	
Superior parietal gyrus	SPG		✓	
Inferior parietal, but supramarginal and angular gyri	IPL		✓	
Precuneus	PCUN		✓	
Superior temporal gyrus	STG		✓	
Temporal pole: superior temporal gyrus	TPOsup		✓	
Middle temporal gyrus	MTG		✓	
Temporal pole: middle temporal gyrus	TPOmid		✓	
Inferior temporal gyrus	ITG		✓	

* where '✓' indicates selection of the respective pair of symmetrical ROIs.

much head motion are excluded. The normalized brain images are warped into AAL atlas to obtain 90 ROIs as nodes. 42 (21 pairs of) ROIs that are known to be related to AD are selected by following [15] in our experiment, and the mean rs-fMRI signal within each ROI is extracted as the features. As on ADHD-200 data set, these selected 21 pairs of symmetrical ROIs are illustrated in Fig. 2(b) with each pair sharing a distinct color, and the mappings between colors and ROI names are shown in Table I.

The ROI mean time series from the two data sets are then band-pass filtered with frequency interval ($0.025 \leq f \leq 0.100$ Hz) to obtain the most discriminative frequency band and this frequency interval is further decomposed into five equal-length spectral by following [49] since it has been demonstrated that such different sub-bands provide extra frequency-specific clues of BOLD signals, enabling a more frequency-specific analysis of the regional mean time series.

B. Experimental Settings

The input data have six channels formed by concatenating the ROI mean time series within the frequency interval of $0.025 \leq f \leq 0.100$ Hz and the signals from the five decomposed equal-length spectrums. As mentioned above, the ROI numbers N of ADHD-200 and ADNI data sets are 24 and 42, respectively. A time frame window of $T = 120$ is randomly sampled from the total time frames as data augmentation

during the training stage. Four stacked Transformer encoder layers are used with $H = 4$ heads and the dimensions D set as 32, 64, 64, 128, respectively, in the four layers. The diffusion order S in Eq. (8) is set as three. Regarding the training/test splits, the predefined partition for ADHD-200 is used. On ADNI data set, 80% of the samples are randomly selected for training and the remaining 20% for test. And this random partition is repeated ten times to obtain reliable statistics. During model training, the initial learning rate is set as 0.1 and divided by 10 in 60 and 80 epochs with a total number of 100 epochs and batch size of 10. The stochastic gradient descent (SGD) is used as the optimizer with cross entropy as the loss function to train the model.

C. Experimental Results on ADHD-200 Data Set

1) *Methods for Comparison*: The state-of-the-art methods are compared with our proposed method to verify its effectiveness. The competing methods are briefly introduced as follows.

- **MKL** method [50]: The multiple kernel learning (MKL) method extracts various features from both structural magnetic resonance imaging (sMRI) and functional magnetic resonance imaging (fMRI). These features are integrated by MKL and classified with support vector machines (SVM).
- **AGDM** method [51]: In the attributed graph distance measure (AGDM) method, FBNs are constructed by defining the nodes as the clusters of highly active voxels

TABLE II
SUMMARY OF THE ADHD-200 DATA SET

Training Data Set								
	KKI	NeuroIMAGE	NYU	OHSU	PKU	UPittsburgh	WashU	Total
#sub	83	48	216	79	194	89	59	768
Control	61	23	98	42	116	89	59	488
ADHD	22	25	118	37	78	0	0	280
Age	8-13	11-22	7-18	7-12	8-17	10-20	7-22	7-22
Male	46	31	140	43	144	46	32	482
Female	37	17	76	36	50	43	27	286
Test Data Set								
	KKI	NeuroIMAGE	NYU	OHSU	PKU	UPittsburgh		Total
#sub	11	25	41	34	51	9		171
Control	8	14	12	28	27	5		94
ADHD	3	11	29	6	24	4		77
Age	8-12	13-26	7-17	7-12	8-15	14-17		7-26
Male	10	12	28	17	32	7		106
Female	1	13	13	17	19	2		65

* where '#sub' denotes the number of subjects.

and edges as the correlations. A local node descriptor comprising of a set of attributes is computed for classification by SVM.

- **Social-net** method [52]: It constructs correlation-based social network from rs-fMRI and extracts features of assortative mixing and synchronization in addition to the traditional network features for classification with SVM.
- **3D CNN** method [53]: This is a deep learning-based method via 3D convolutional neural networks (CNN). It first extracts various 3D low-level features from sMRI and fMRI data. Then a multi-modality CNN architecture is designed to combine the fMRI and sMRI features.
- **DBN** method [54]: The deep belief network (DBN) method combines DBN and Bayesian network into a integrated model for dimension reduction and feature extraction. Then the extracted features are classified by SVM.
- **BNS** method [55]: In the brain's network structure (BNS) method, FBNS are constructed by computing the pairwise correlation of brain voxels' activities over the time frames. Various network features are extracted from each voxel in the network. The network features of all the voxels in a brain are concatenated and serve as the feature vector to train a principal component analysis-linear discriminant analysis (PCA-LDA) based classifier for classification.
- **EM-MI** method [45]: The expectation-maximization multi-instance learning (EM-MI) method proposes a short-time classification technology to analyze the fMRI data by evaluating the correlation degree between each fMRI segment and ADHD. The degree is then used as feature for classifier training with multiple-instance learning.
- **4D CNN** method [20]: In this work, a spatio-temporal 4D CNN model is proposed to analyze rs-fMRI data. The method is able to calculate granularity at a coarse level by stacking layers. With rs-fMRI as 3D time-series frames, several models of spatial and temporal granular computing and fusion are exploited, including feature pooling, long short-term memory (LSTM) and spatio-temporal convolution for classification.

- **SASNI** method [56]: The subject-adaptive SICE network integration (SASNI) method constructs FBNS with sparse inverse covariance estimation (SICE) and proposes a subject-adaptive method to integrate multiple SICE networks as a unified representation for classification. The integration weights are learned adaptively for each subject in order to endow the method with the flexibility in dealing with subject variations. Furthermore, to respect the manifold geometry of SICE networks, Stein kernel is employed to embed the manifold structure into a kernel-induced feature space, which allows a linear integration of SICE networks to be designed.
- **HO-FCN** method [57]: This method constructs high-order functional connectivity networks (HO-FCN) for MCI classification. Specifically, a sliding window approach is utilized first to partition the entire rs-fMRI time series into multiple segments. Then, multiple low-order functional connectivity networks are constructed to further build the correlation time series for each brain region pairs. Finally, the correlation between correlation time series can be calculated, from which graph features can be derived for SVM classification.
- **ST-GCN** method [28]: This method proposes to formulate FBNS within the context of spatio-temporal graphs. A spatio-temporal graph convolutional network (ST-GCN) is constructed and trained on short sub-sequences of the BOLD time series to model the non-stationary nature of functional connectivity. At the same time, the model optimizes the importance of graph edges within ST-GCN to gain discriminative information of the functional connectivities.

2) Performance Evaluation: Diagnostic performance of the competing methods mentioned above and the proposed method on ADHD-200 data set is summarized in Table III, where the best results are highlighted in bold. As can be seen, the upper portion of this table quotes the state-of-the-art results in the literature, while the lower portion lists the results of the methods implemented by this work, including two variants of our proposed method as introduced below. All the methods

TABLE III
COMPARISON OF THE CLASSIFICATION PERFORMANCE (IN%) BETWEEN THE-STATE-OF-THE-ART METHODS
AND THE PROPOSED KERNEL TRANSFORMER ON ADHD-200 DATA SET

Test Data Set							
Imaging site	PKU	NYU	OHSU	NeuroIMAGE	KKI	UPittsburgh	Overall
Number of subjects	51	41	34	25	11	9	171
Quoted results from literature							
MKL (2012) [50]	—	—	—	—	—	—	61.5
AGDM (2014) [51]	58.8	—	82.4	48.0	54.6	—	62.8
Social-net (2014) [52]	—	—	—	—	—	—	63.5
DBN (2015) [54]	66.3	64.7	—	—	59	—	66.3
CNN+LSTM (2019) [20]	—	—	—	—	—	—	68.8
3D CNN (2017) [53]	63.0	70.5	—	—	72.8	—	69.2
BNS (2012) [55]	62.7	70.7	73.5	72.0	72.7	77.8	69.6
EM-MI (2020) [45]	70.6	63.4	—	—	81.8	—	70.4
4D CNN (2019) [20]	—	—	—	—	—	—	71.3
SASNI (2017) [56]	74.5	70.7	79.4	72.0	63.6	55.6	72.5
Results obtained by this paper							
HO-FCN [57]	56.9	70.7	82.4	76.0	81.8	66.7	70.2
ST-GCN (2020) [28]	62.0	70.7	76.5	72.0	81.8	66.7	70.2
Transformer [29]	66.7	78.0	82.4	64.0	81.8	66.7	73.1
K-Transformer (proposed)	70.6	80.4	85.3	68.0	81.8	77.8	75.6
KD-Transformer (proposed)	70.6	82.9	85.3	72.0	90.9	77.8	78.4

* where ‘—’ indicates that the performance on the respective imaging site is not reported.

involved use the standard training/test sets predefined by the data set. In order to conduct a comprehensive comparison and analysis, both of the overall results on the whole test set and the detailed evaluations available on each imaging site are collected and reported. The comparison methods in each portion are listed in ascending order with respect to their overall performance. Transformer [29] refers to the conventional Transformer method using linear attention mechanism. K-Transformer denotes our proposed kernel attention Transformer, whose attention module is carried out by an RBF kernel function, as introduced in Section II-C. Note that RBF kernel is recommended as a general case if no known prior knowledge is assumed. In our work, the parameter β of the RBF kernel is automatically learned to fit the task of brain disorder classification. KD-Transformer indicates our proposed complete solution by applying diffusion on top of the kernel attention mechanism, as in Section II-D. As seen in the table, Transformer [29] achieves an promising accuracy of 73.1%. It is higher than the accuracies of all the traditional machine learning methods based on hand-crafted features, including MKL [50] (61.5%), AGDM [51] (62.8%), Social-net [52] (63.5%), BNS [55] (69.6%), HO-FCN [57] (70.2%), EM-MI [45] (70.4%), SASNI [56] (72.5%). At the same time, it also outperforms various deep learning methods, including Deep Bayesian Network (DBN) method [54] (66.3%), CNN and RNN based methods CNN+LSTM [20] (68.8%), 3D CNN [53] (69.2%), 4D CNN [20] (71.3%), and GCN based ST-GCN [28] (70.2%). This verifies the encouraging effectiveness of the attention-based Transformer [29] method in modeling functional dependencies between brain ROIs. With the proposed kernel attention mechanism, K-Transformer further improves the accuracy to 75.6%, indicting the advantage of modeling non-linear relationships between brain regions. The proposed KD-Transformer further boosts the diagnostic accuracy to 78.4%, with an improvement of 5.9 percentage points over the state-of-the-art method SASNI [56] and 5.3 percentage points over the conventional Transformer [29]. Regarding

the performance on each imaging site, KD-Transformer achieves the best accuracy on NYU, OHSU, NeuroIMAGE, KKI and UPittsburgh sites, and its performance is only inferior to SASNI [56] on the PKU site.

This result is encouraging, and the improvements of the proposed methods are significant, indicating the efficacy of the Transformer framework and also the proposed diffusion kernel attention mechanism.

D. Experimental Results on ADNI Data Set

The ADNI data set is also used to evaluate the performance of the proposed diffusion kernel attention Transformer. Different from ADHD-200 data set, the data pre-processing and training/test partitioning on this data set are conducted by this work since there is no publicly released pre-processed rs-fMRI data or standard training/test split as on ADHD-200 data set. In this case, in order to conduct a fair comparison, we implement various competing methods and perform them with identical processed rs-fMRI data and training/test partitions. The competing methods on this data set are listed as follows.

1) Methods for Comparison:

- **LCC** method [17]: The local clustering coefficient (LCC) method extracts LCC from functional brain networks as feature representation of the topological properties of each brain node. Then the feature feeds into SVM for classification.
- **ST-GCN** method [28]: Refer to Section III-C.1.
- **HO-FCN** method [57]: Refer to Section III-C.1.
- **Compact Representation** method [19]: This method studies the manifold properties of functional brain networks constructed with SICE method and employs manifold-based similarity measures and kernel-based principal component analysis (PCA) to extract leading principal connectivity components for the compact representation of brain networks for classification.

TABLE IV

COMPARISON OF THE CLASSIFICATION PERFORMANCE (IN%)
BETWEEN THE STATE-OF-THE-ART METHODS AND THE PROPOSED
KERNEL TRANSFORMER ON ADNI DATA SET

Method	Accuracy	p-value
LCC [17]	68.0	
ST-GCN [28]	69.3	
HO-FCN [57]	71.8	
Compact Representation [19]	72.0	
Transformer [29]	73.3	—
K-Transformer (proposed)	77.3	0.0051
KD-Transformer (proposed)	80.0	0.0011

* where '—' indicates the baseline to calculate p -values.

2) Performance Evaluation: Table IV reports the classification accuracy of the competing methods and the proposed methods averaged over 10 training/test splits. 90 ROIs are used in traditional machine learning methods, i.e., LCC [17] and Compact Representation [19], since they admit better performance while 42 ROIs are used in other deep-learning based methods. As seen, the conventional Transformer [29] method achieves an accuracy of 73.3% and it beats LCC [17] (68.0%), HO-FCN method [57] (71.8%) and Compact Representation [19] (72.0%), and recent advanced deep learning method ST-GCN [28] (69.3%), with considerable margins. With the proposed kernel attention mechanism and diffusion process, the accuracy is further improved to 77.3% by K-Transformer and 80% by KD-Transformer, obtaining improvements of 4.0 and 6.7 percentage points, respectively, over Transformer [29]. Besides classification accuracy, the p -value obtained by paired-samples t -test between the proposed method and Transformer [29] is used to evaluate the significance of improvement (where p -value ≤ 0.05 is used). As seen, the p -values of the proposed K-Transformer and KD-Transformer over Transformer [29] are 0.0051 and 0.0011, respectively, which indicate the statistical significance of the improvements.

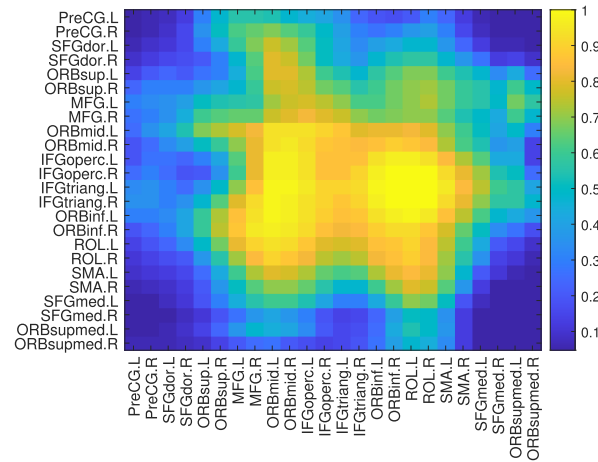
This result is consistent with the evaluation on ADHD-200 data set in the above section. They jointly verify that the proposed diffusion kernel attention mechanism is effective in functional brain network construction and analysis, admitting promising potentials in brain disorder classification.

IV. DISCUSSION

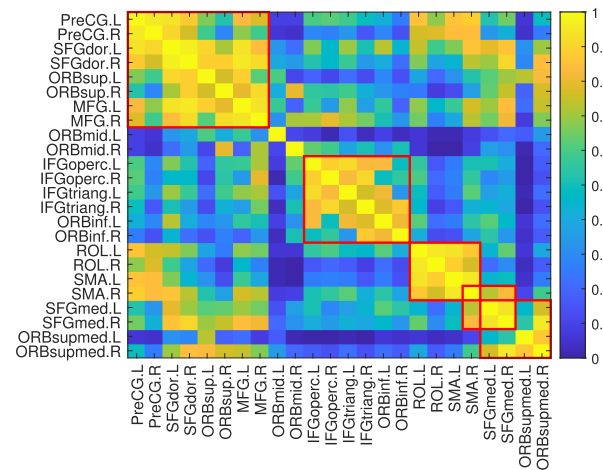
A. Visualization of the Learned Attention Maps

The last section has evaluated the promising performance of the proposed method. In order to intuitively present what the proposed diffusion kernel attention mechanism learns and gain insights on the difference between the conventional Transformer and the proposed method, some example attention maps of these two methods are visualized in this section.

Firstly, a comparison is conducted between the proposed method and the competing method with the same backbone network but without the proposed attention mechanism. Fig. 3 plots the example attention maps on ADHD-200 data set learned by the conventional linear Transformer and the proposed kernel Transformer, which correspond to A and the kernel variant K , respectively. Note that both A and K can be



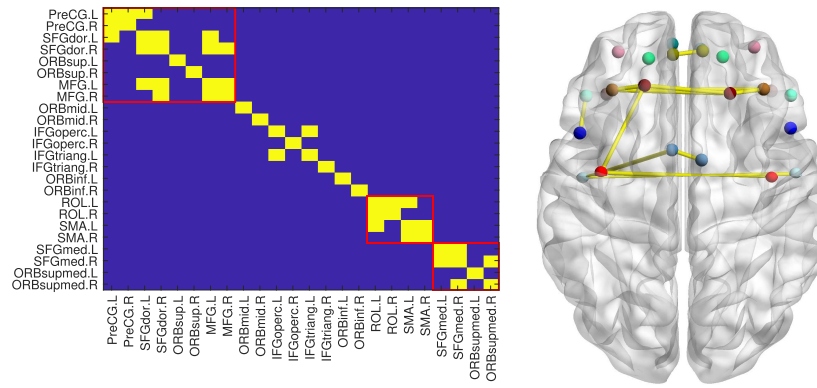
(a) Example linear attention map



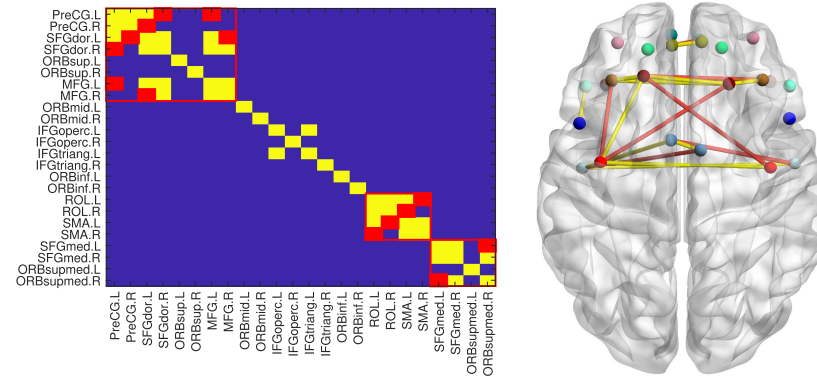
(b) Example kernel attention map

Fig. 3. Example functional brain networks (attention maps) on ADHD-200 data set learned by the conventional linear Transformer in (a) and the proposed kernel Transformer in (b). The tick-labels of the x- and y-axis in each figure show the region of interest (ROI) acronyms, which are defined in Table I and the suffix 'L' and the suffix 'R' denote the left and the right hemispheres, respectively. In comparison with (a) where only dominant and concentrated connection patterns are learned, the proposed kernel attention in (b) learns richer block-wise connection patterns, demonstrating the modular organization of the human brain. These richer and reasonable patterns obtained by the proposed kernel attention method could better reflect the underlying human brain connections, and more informative clues could be extracted based on these patterns to benefit the brain disorder classification.

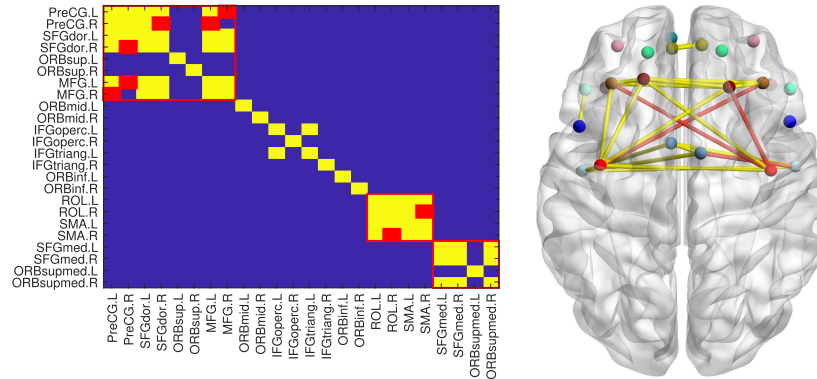
interpreted as functional brain networks (FBN) constructed in the Transformer architecture. In each plot, the x-axis shows the ROI index while the y-axis provides the ROI names, and the colorbar indicates the connection strength. As can be seen, linear Transformer learns FBN with only dominant and concentrated connection patterns as in in Fig. 3(a), while the proposed kernel attention in Fig. 3(b) learns richer block-wise connection patterns, demonstrating the modular organization of the human brain [58]. Specifically, the connection patterns are characterized by dense local clustering or cliquishness of connections between neighboring nodes within multiple brain node groups, as highlighted by the red rectangles. In contrast, there exist relatively fewer long-range connec-



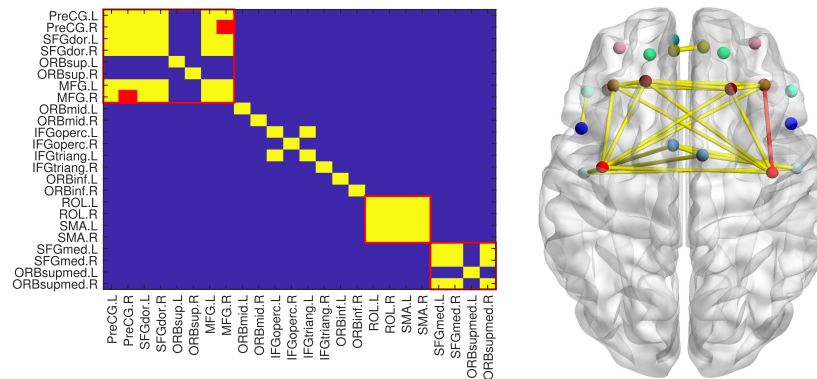
(a) Kernel attention map without diffusion



(b) Kernel attention map with the 1st order diffusion



(c) Kernel attention map with the 2nd order diffusion



(d) Kernel attention map with the 3rd order diffusion

Fig. 4. The effects of diffusion process on kernel attention in transformer. (a) shows a binary kernel attention map in matrix and brain views, and (b), (c) and (d) illustrate three different orders of diffusion processes conducted on the binary attention map, respectively. The newly added attention entries by the current order of diffusion process are highlighted as red in both matrix and brain views. The tick-labels of the x- and y-axis in each matrix view show the region of interest (ROI) acronyms, which are defined in Table I and the suffix 'L' and the suffix 'R' denote the left and the right hemispheres, respectively. Each pair of symmetrical ROIs in each brain view shares a distinct color and the corresponding ROI names of different colors are also provided in Table I.

tions across groups. Such block-wise connection patterns have been observed in many studies and have been verified as an attractive organization of brain networks since they can support both specialized and integrated information processing in an economical manner [59]. In short, these patterns obtained by the proposed kernel attention are reasonable and they reveal richer non-linear modular dependencies between brain ROIs which can hardly be fully modeled by the linear attention mechanism. These richer and reasonable patterns obtained by the proposed kernel attention method could better reflect the underlying human brain connections, and more informative clues could be extracted based on these patterns to benefit the brain disorder classification.

Moreover, the effects of diffusion process on kernel attention in Transformer are illustrated in Fig. 4 by using two views, i.e., the matrix view and the brain view. In order to better demonstrate the effects, we only keep the strong connections in the kernel attention map by applying a threshold of 0.8, and the resulting attention map is binarized, as shown in Fig. 4(a). Then three orders ($S = 3$) of diffusion process are applied on the binary attention map, which are illustrated in Fig. 4(b), Fig. 4(c) and Fig. 4(d), respectively. Each time, the newly added attention entries by the current order of diffusion process are highlighted as red in both matrix and brain views. The illustration can be interpreted from both the overall and the detailed perspectives respectively.

Firstly, from an overall perspective, with the increase of the diffusion orders, the range of connections is getting wider. This enables wider interactions between indirectly connected brain regions, providing more complex activity patterns for brain function analysis. Specifically, as seen in Fig. 4(a) when only kernel attention is applied without diffusion process, the modular structure, which is highlighted by red rectangles, can be roughly observed but is vague. When the first order ($S = 1$) diffusion process is applied in Fig. 4(b), many new (marked as red) connections are introduced. More interestingly, all the newly added red entries fall in the three highlighted modules, demonstrating that the diffusion process tends to increase the connections within modules and therefore enhance the modular organization of the FBNs constructed by the kernel attention. Higher orders of diffusion in Fig. 4(c) ($S = 2$) and Fig. 4(d) ($S = 3$) also follow this trend and further reinforce the modular structure. Also, with the increase of the order S , the number of newly introduced red attention entries decreases when the connections within modules have been very dense. As introduced above, the modular structure is a typical characteristic of the human brain organization [58], so the enhanced modular structure by the diffusion process could make the resulting FBNs better aligned with the human brain organization and in turn benefit the feature extraction and brain disorder classification.

Secondly, from a detailed perspective by looking into the specific connections, the final connections obtained by the proposed diffusion kernel attentions in Fig. 4(d) mainly include three blocks of connections: 1) connections among Precentral gyrus (PreCG in short), Superior frontal gyrus, dorsolateral (SFGdor) and Middle frontal gyrus (MFG); 2) connections between Rolandic operculum (ROL) and Supplementary motor

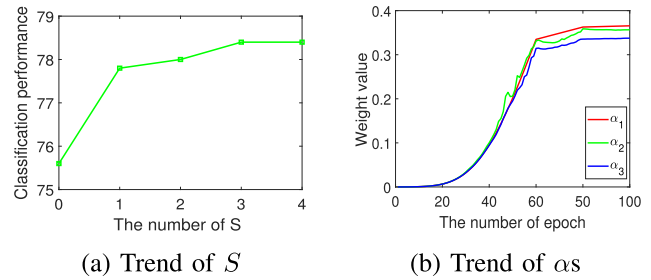


Fig. 5. Ablation studies on the order S and combination weights α_s . (a) demonstrates that applying diffusion on the kernel attention (when $S > 0$) consistently outperforms the sole kernel attention (when $S = 0$), and larger S leads to better classification performance until S reaches 3. As seen in (b), the trends of different α_s are similar, i.e., there is a rapid increase at the beginning of the training process and it is followed by a slow growth, and they finally converge to certain values. The convergent values of α_1 and α_2 are similar, and they are larger than α_3 .

area (SMA); and 3) connections between Superior frontal gyrus, medial (SFGmed) and Superior frontal gyrus, medial orbital (ORBsupmed). As can be seen from the corresponding brain view on the right, all the ROIs in each group are spatial neighbors. This is reasonable since structural neighbors are more likely to be functionally linked. Besides, the above ROIs in these three blocks have been found highly related to ADHD in the literature. For example, in comparison with healthy controls, [60], [61] identified significantly smaller grey matter volume in PreCG and MFG of ADHD patients, and [62] found decreased regional homogeneity (ReHo) in PreCG and SMA accordingly while [63] observed decreased activation in PreCG, SFGdor, MFG, SFGmed and SMA. This verifies that the proposed diffusion kernel attention method effectively identifies the disease-related ROIs to guide further analysis, i.e., ROI feature extraction and brain disorder classification.

B. Ablation Studies on S and α_s

Ablation studies on the order S and combination weights α_s are conducted to verify their impacts on the classification performance of the proposed method. The study on S is presented in Fig. 5(a) while the study on α_s is reported in Fig. 5(b).

Fig. 5(a) demonstrates that applying diffusion on the kernel attention (when $S > 0$) consistently outperforms the sole kernel attention (when $S = 0$), and larger S leads to better classification performance until S reaches 3. So, $S = 3$ is used in our paper. This choice is also consistent with Fig. 4 which shows that with the increase of the order S , the number of newly introduced red attention entries decreases while the 3rd order diffusion introduces only one red entry. It can be expected that higher order diffusion will have minor effects on the connection patterns.

Regarding the combination weights α_s , as seen in Fig. 5(b), the trends of different α_s are similar, i.e., there is a rapid increase at the beginning of the training process and it is followed by a slow growth, and they finally converge to certain values. The convergent values of α_1 and α_2 are similar, and they are larger than α_3 . This reflects lower orders of diffusion are more important.

V. CONCLUSION

In this paper, inspired by Transformer, we propose a diffusion kernel attention network for brain disorder classification by adaptively integrating the FBN construction and classification into a unified model and training it in an end-to-end manner. Specifically, after thorough analysis of the limitations of the vanilla Transformer architecture in brain disorder classification, the proposed method innovatively adapts the conventional Transformer from two perspectives:

- Kernel attention is proposed as a substitute of the original linear attention module. This not only significantly reduces the number of parameters to train and thus alleviate the issue of small samples, but also introduces non-linear attention mechanism to model complex interactions between brain regions;
- Diffusion is further employed to the kernel attention module to incorporate wider interactions among indirectly connected brain regions as well.

The experimental results on ADHD and AD classification demonstrates superior performance of the proposed method to the competing methods.

However, there still exists several limitations that will be explored in our future work. Firstly, only RBF kernel is employed in the proposed method, while discovering better kernels for temporal signal is worth exploration. Secondly, how to automatically determine the orders of diffusion process through optimization is also a critical research issue. Last but not least, the generability of the proposed method will be further explored. The proposed method can be easily applied to many applications considering that 1) the proposed diffusion kernel attention is independent of specific Transformer architecture and it can be readily used as a general tool to replace the linear attention easily in many attention based models; and 2) the kernel function can be flexibly designed to fit different data format or incorporate domain knowledge.

REFERENCES

- [1] Y. Fan *et al.*, "Multivariate examination of brain abnormality using both structural and functional MRI," *NeuroImage*, vol. 36, no. 4, pp. 1189–1199, Jul. 2007.
- [2] J. Richiardi *et al.*, "Classifying minimally disabled multiple sclerosis patients from resting state functional connectivity," *NeuroImage*, vol. 62, no. 3, pp. 2021–2033, Sep. 2012.
- [3] M. P. van den Heuvel and H. E. H. Pol, "Exploring the brain network: A review on resting-state fMRI functional connectivity," *Eur. Neuropsychopharmacol.*, vol. 20, no. 8, pp. 519–534, 2010.
- [4] G. Chen *et al.*, "Classification of Alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging," *Radiology*, vol. 259, no. 1, p. 213, 2011.
- [5] X. Chen, H. Zhang, L. Zhang, C. Shen, S.-W. Lee, and D. Shen, "Extraction of dynamic functional connectivity from brain grey matter and white matter for MCI classification," *Hum. Brain Mapping*, vol. 38, no. 10, pp. 5019–5034, 2017.
- [6] H. Luan, F. Qi, Z. Xue, L. Chen, and D. Shen, "Multimodality image registration by maximization of quantitative–qualitative measure of mutual information," *Pattern Recognit.*, vol. 41, no. 1, pp. 285–298, 2008.
- [7] L. Farràs-Permanyer, J. Guàrdia-Olmos, and M. Peró-Cebollero, "Mild cognitive impairment and fMRI studies of brain functional connectivity: The state of the art," *Frontiers Psychol.*, vol. 6, p. 1095, Aug. 2015.
- [8] T.-E. Kam, H. Zhang, Z. Jiao, and D. Shen, "Deep learning of static and dynamic brain functional networks for early MCI detection," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 478–487, Feb. 2020.
- [9] H. Jia, G. Wu, Q. Wang, and D. Shen, "ABSORB: Atlas building by self-organized registration and bundling," *NeuroImage*, vol. 51, no. 3, pp. 1057–1070, Jul. 2010.
- [10] H. Jia, P.-T. Yap, and D. Shen, "Iterative multi-atlas-based multi-image segmentation with tree-based registration," *NeuroImage*, vol. 59, no. 1, pp. 422–430, 2012.
- [11] B. Jie, M. Liu, and D. Shen, "Integration of temporal and spatial properties of dynamic connectivity networks for automatic diagnosis of brain disease," *Med. Image Anal.*, vol. 47, pp. 81–94, Jul. 2018.
- [12] B. C. Munsell *et al.*, "Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data," *NeuroImage*, vol. 118, pp. 219–230, Sep. 2015.
- [13] S. M. Smith *et al.*, "Network modelling methods for FMRI," *NeuroImage*, vol. 54, no. 2, pp. 875–891, 2011.
- [14] N. Tzourio-Mazoyer *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, Jan. 2002.
- [15] S. Huang *et al.*, "Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation," *NeuroImage*, vol. 50, no. 3, pp. 935–949, 2010.
- [16] C.-Y. Wee, P.-T. Yap, D. Zhang, L. Wang, and D. Shen, "Constrained sparse functional connectivity networks for MCI classification," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent*, in Lecture Notes in Computer Science, vol. 7511. Nice, France, 2012, pp. 212–219.
- [17] M. Kaiser, "A tutorial in connectome analysis: Topological and spatial features of brain networks," *NeuroImage*, vol. 57, no. 3, pp. 892–907, 2011.
- [18] N. Leonardi *et al.*, "Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest," *Neuroimage*, vol. 83, pp. 937–950, Dec. 2013.
- [19] J. Zhang, L. Zhou, L. Wang, and W. Li, "Functional brain network classification with compact representation of SICE matrices," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 6, pp. 1623–1634, Jun. 2015.
- [20] Z. Mao *et al.*, "Spatio-temporal deep learning method for ADHD fMRI classification," *Inf. Sci.*, vol. 499, pp. 1–11, Oct. 2019.
- [21] J. Huang, L. Zhou, L. Wang, and D. Zhang, "Attention-diffusion-bilinear neural network for brain network analysis," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2541–2552, Jul. 2020.
- [22] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1195–1206, May 2019.
- [23] E. Jeon, E. Kang, J. Lee, J. Lee, T.-E. Kam, and H.-I. Suk, "Enriched representation learning in resting-state fMRI for early MCI diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent*, vol. 12267. Lima, Peru, 2020, pp. 397–406.
- [24] Q. Dong, J. Qiang, J. Lv, X. Li, T. Liu, and Q. Li, "Spatiotemporal attention autoencoder (STAAE) for ADHD classification," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent* in Lecture Notes in Computer Science, vol. 12267. Lima, Peru, 2020, pp. 508–517.
- [25] X. Hu *et al.*, "Latent source mining in FMRI via restricted Boltzmann machine," *Hum. Brain Mapping*, vol. 39, no. 6, pp. 2368–2380, Jun. 2018.
- [26] R. D. Hjelm, V. D. Calhoun, R. Salakhutdinov, E. A. Allen, T. Adali, and S. M. Plis, "Restricted Boltzmann machines for neuroimaging: An application in identifying intrinsic networks," *Neuroimage*, vol. 96, no. 8, pp. 245–260, 2014.
- [27] L. Zhang *et al.*, "Deep fusion of brain structure-function in mild cognitive impairment," *Med. Image Anal.*, vol. 72, pp. 1–17, Aug. 2021.
- [28] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, "Spatio-temporal graph convolution for resting-state fMRI analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent* in Lecture Notes in Computer Science, vol. 12267. Lima, Peru, 2020, pp. 528–538.
- [29] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [30] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [31] M. Chen *et al.*, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.
- [32] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

- [33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
- [34] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
- [35] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive Transformers with linear attention," in *Proc. Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 5156–5165.
- [36] K. Han *et al.*, "A survey on visual transformer," 2020, *arXiv:2111.06091*.
- [37] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, pp. 1–38, Dec. 2021, doi: [10.1145/3505244](https://doi.org/10.1145/3505244).
- [38] Y.-H.-H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Transformer dissection: A unified understanding for Transformer's attention via the lens of kernel," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4344–4353.
- [39] S. M. Smith, "The future of fMRI connectivity," *Neuroimage*, vol. 62, no. 2, pp. 1257–1266, 2012.
- [40] R. Li, J. Su, C. Duan, and S. Zheng, "Linear attention mechanism: An efficient attention for semantic segmentation," 2020, *arXiv:2007.14902*.
- [41] G. Rosenthal *et al.*, "Mapping higher-order relations between brain structure and function with embedded vector representations of connectomes," *Nature Commun.*, vol. 9, no. 1, pp. 1–12, Dec. 2018.
- [42] L. Zhang, L. Wang, and D. Zhu, "Recovering brain structural connectivity from functional connectivity via multi-GCN based generative adversarial network," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, vol. 12267, Lima, Peru, 2020, pp. 53–61.
- [43] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," in *Proc. Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2002, pp. 315–322.
- [44] J. D. Noh and H. Rieger, "Random walks on complex networks," *Phys. Rev. Lett.*, vol. 92, no. 11, Mar. 2004, Art. no. 118701.
- [45] 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.
- [46] B. Jie, D. Zhang, B. Cheng, and D. Shen, "Manifold regularized multitask feature learning for multimodality disease classification," *Hum. Brain Mapping*, vol. 36, no. 2, pp. 489–507, 2015.
- [47] H.-I. Suk, S.-W. Lee, and D. Shen, "Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis," *Brain Struct. Function*, vol. 221, no. 5, pp. 2569–2587, 2016.
- [48] C.-G. Yan and Y.-F. Zang, "DPARSF: A MATLAB toolbox for 'pipeline' data analysis of resting-state fMRI," *Frontiers Syst. Neurosci.*, vol. 4, pp. 1–7, May 2010.
- [49] C.-Y. Wee *et al.*, "Resting-state multi-spectrum functional connectivity networks for identification of MCI patients," *PLoS ONE*, vol. 7, no. 5, pp. 1–11, May 2012, doi: [10.1371/journal.pone.0037828](https://doi.org/10.1371/journal.pone.0037828).
- [50] D. Dai, J. Wang, J. Hua, and H. He, "Classification of ADHD children through multimodal magnetic resonance imaging," *Frontiers Syst. Neurosci.*, vol. 6, p. 63, Sep. 2012.
- [51] S. Dey, A. R. Rao, and M. Shah, "Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects," *Frontiers Neural Circuits*, vol. 8, pp. 1–11, Jun. 2014.
- [52] X. Guo, X. An, D. Kuang, Y. Zhao, and L. He, "ADHD-200 classification based on social network method," in *Proc. Int. Conf. Intell. Comput.*, vol. 8590, Taiyuan, China, 2014, pp. 233–240.
- [53] 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.
- [54] A. J. Hao, C. H. Yin, and B. L. He, "Discrimination of ADHD children based on deep Bayesian network," in *Proc. IET Int. Conf. Biomed. Image Signal Process. (ICBISP)*, 2015, pp. 1–6.
- [55] S. Dey, A. R. Rao, and M. Shah, "Exploiting the brain's network structure in identifying ADHD subjects," *Frontiers Syst. Neurosci.*, vol. 6, pp. 1–13, Nov. 2012.
- [56] J. Zhang, L. Zhou, and L. Wang, "Subject-adaptive integration of multiple SICE brain networks with different sparsity," *Pattern Recognit.*, vol. 63, pp. 642–652, Mar. 2017.
- [57] X. Chen *et al.*, "High-order resting-state functional connectivity network for MCI classification," *Hum. Brain Mapping*, vol. 37, no. 9, pp. 3282–3296, 2016.
- [58] X. Yang, L. Shi, M. Daianu, H. Tong, Q. Liu, and P. Thompson, "Blockwise human brain network visual comparison using NodeTrix representation," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 181–190, Jan. 2017.
- [59] D. S. Bassett and E. T. Bullmore, "Small-world brain networks revisited," *Neuroscientist*, vol. 23, no. 5, pp. 499–516, Sep. 2017.
- [60] J. Bralten *et al.*, "Voxel-based morphometry analysis reveals frontal brain differences in participants with ADHD and their unaffected siblings," *J. Psychiatry Neurosci.*, vol. 41, no. 4, pp. 272–279, Jul. 2016.
- [61] L.-J. Wang *et al.*, "Gray matter vol. and, microRNA levels in patients with attention-deficit/hyperactivity disorder," *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 270, no. 8, pp. 1037–1045, 2020.
- [62] C. Y. Shang, H. Y. Lin, W. Y. Tseng, and S. S. Gau, "A haplotype of the dopamine transporter gene modulates regional homogeneity, gray matter volume, and visual memory in children with attention-deficit/hyperactivity disorder," *Psychol. Med.*, vol. 48, no. 15, pp. 2530–2540, Nov. 2018.
- [63] D. Lei *et al.*, "Functional MRI reveals different response inhibition between adults and children with ADHD," *Neuropsychology*, vol. 29, no. 6, p. 874, 2015.