NeuroImage

Gray Matters: ViT-GAN Framework for Identifying Schizophrenia Biomarkers Linking Structural MRI and Functional Connectivity

--Manuscript Draft--

Manuscript Number:	NIMG-24-53R2
Article Type:	Full Length Article
Section/Category:	Computational modelling and analysis
Corresponding Author:	Yuda Bi, Ph.D Georgia State University, Center for Translational Research in Neuroimaging and Data Science (TReNDS) Atlanta, GA UNITED STATES
First Author:	Yuda Bi, Ph.D
Order of Authors:	Yuda Bi, Ph.D
	Anees Abrol
	Sihan Jia
	Jing Sui
	Vince Calhoun
Abstract:	Brain disorders are often associated with changes in brain structure and function, where functional changes may be due to underlying structural variations. Gray matter (GM) volume segmentation from 3D structural MRI offers vital structural information for brain disorders like schizophrenia, as it encompasses essential brain tissues such as neuronal cell bodies, dendrites, and synapses, which are crucial for neural signal processing and transmission; changes in GM volume can thus indicate alterations in these tissues, reflecting underlying pathological conditions. In addition, the use of the ICA algorithm to transform high-dimensional fMRI data into functional network connectivity (FNC) matrices serves as an effective carrier of functional information. In our study, we introduce a new generative deep learning architecture, the conditional efficient vision transformer generative adversarial network (cEViT-GAN), which adeptly generates FNC matrices conditioned on GM to facilitate the exploration of potential connections between brain structure and function. We developed a new, lightweight self-attention mechanism for our ViT-based generator, enhancing the generation of refined attention maps critical for identifying structural biomarkers based on GM. Our approach not only generates high quality FNC matrices with a Pearson correlation of 0.74 compared to real FNC data, but also uses attention map technology to identify potential biomarkers in GM structure that could lead to functional abnormalities in schizophrenia patients. Visualization experiments within our study have highlighted these structural biomarkers, including the medial prefrontal cortex (mPFC), dorsolateral prefrontal cortex (DL-PFC), and cerebellum. In addition, through cross-domain analysis comparing generated and real FNC matrices, we have identified functional connections with the highest correlations to structural information, further validating the structure-function connections. This comprehensive analysis helps to understand the intricate relationship between brain structure and its functional manifestations, providing a more refined insight into the neurobiological research of schizophrenia.
Suggested Reviewers:	Tianming Liu tliu@uga.edu
	Dinggang Shen dinggang.shen@gmail.com

Dear Editorial Board:

We are submitting our manuscript entitled "Gray Matters: An Efficient Vision Transformer GAN Framework for Predicting Functional Network Connectivity Biomarkers from Brain Structure" for consideration in your esteemed journal, NeuroImage. We confirm that this manuscript is original, has not been published previously, and is not under consideration by any other journal.

We have obtained the necessary permissions to republish any figures, tables, or other material that has been previously published. We also disclose that this research has no conflicts of interest to declare.

Yours sincerely,

Yuda Bi
Anees Abrol
Sihan Jia
Jing Sui
Vince Calhoun

Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS),
{ybi3, vcalhoun}@gsu.edu

Highlight

- Our ECViT-GAN is an innovative and effective framework for synthesizing functional connectivity from 3D gray matter data.
- Our approach accurately predicts brain function based on structural gray matter brain scans with amazing precision.
- Our model effectively established a connection between structural and functional brain imaging data through the utilization of generative AI.
- We ascertain the crucial anatomical brain areas that have considerable relevance to the functional aspects of schizophrenia.

Reviewer 4:

What was the framework with which the cevit-gan was implemented? Pytorch, Tensorflow, other?

Response: Our cEViT-GAN was implemented using Pytorch framework.

Which all-reduce mechanisms to distribute the load between GPUs was used? Or was it a parameter server approach?

Response: In our implementation, we utilized PyTorch's built-in `torch.nn.parallel` module to manage multi-GPU training. Specifically, we employed the DataParallel class, which utilizes an all-reduce mechanism to ensure efficient load distribution and gradient synchronization across GPUs. This approach allows us to leverage the collective communication capabilities of the GPUs, without the need for a parameter server architecture.

Gray Matters: ViT-GAN Framework for Identifying Schizophrenia Biomarkers Linking Structural MRI and Functional Connectivity

Yuda Bi*, Anees Abrol*, Sihan Jia*, Jing Sui*, Vince D. Calhoun*
*Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory, Atlanta, GA-30303, USA

Abstract—Brain disorders are often associated with changes in brain structure and function, where functional changes may be due to underlying structural variations. Gray matter (GM) volume segmentation from 3D structural MRI offers vital structural information for brain disorders like schizophrenia, as it encompasses essential brain tissues such as neuronal cell bodies, dendrites, and synapses, which are crucial for neural signal processing and transmission; changes in GM volume can thus indicate alterations in these tissues, reflecting underlying pathological conditions. In addition, the use of the ICA algorithm to transform high-dimensional fMRI data into functional network connectivity (FNC) matrices serves as an effective carrier of functional information. In our study, we introduce a new generative deep learning architecture, the conditional efficient vision transformer generative adversarial network (cEViT-GAN), which adeptly generates FNC matrices conditioned on GM to facilitate the exploration of potential connections between brain structure and function. We developed a new, lightweight self-attention mechanism for our ViT-based generator, enhancing the generation of refined attention maps critical for identifying structural biomarkers based on GM. Our approach not only generates high quality FNC matrices with a Pearson correlation of 0.74 compared to real FNC data, but also uses attention map technology to identify potential biomarkers in GM structure that could lead to functional abnormalities in schizophrenia patients. Visualization experiments within our study have highlighted these structural biomarkers, including the medial prefrontal cortex (mPFC), dorsolateral prefrontal cortex (DL-PFC), and cerebellum. In addition, through cross-domain analysis comparing generated and real FNC matrices, we have identified functional connections with the highest correlations to structural information, further validating the structure-function connections. This comprehensive analysis helps to understand the intricate relationship between brain structure and its functional manifestations, providing a more refined insight into the neurobiological research of schizophrenia.

I. INTRODUCTION

The complex relationship between the brain's structural and functional properties is a critical area of research, especially for understanding brain health and disorders, which significantly impact human quality of life. Schizophrenia is a profoundly serious and complex brain disorder with a wide range of symptoms and manifestations. Research suggests that structural changes in the brain, such as changes in the medial prefrontal cortex and enlargement of the lateral ventricles, may occur in the early stages of schizophrenia [39]. Whether these structural changes are causative factors for the functional abnormalities observed in patients with schizophrenia remains a topic worthy of investigation. Structural MRI (sMRI) and functional MRI (fMRI) are essential

neuroimaging tools that offer unique insights into the brain's physiology in health and disease [1], [2], [3]. sMRI provides high-resolution images crucial for detecting morphological changes in gray matter (GM), such as cortical atrophy and hippocampal shrinkage, which are key indicators of neurological conditions like Alzheimer's disease [4], [5]. In contrast, fMRI captures brain activity patterns, identifying regions active during specific tasks or states, and is vital for studying disorders like schizophrenia or autism where these patterns are often disrupted [6]. The interaction between the brain's structural changes and functional abnormalities is intricate and bidirectional. Structural alterations can lead to functional impairments, affecting cognitive processes and behavior. Despite recognition of this connection, it remains not fully understood due to individual differences and the influences of genetic and environmental factors [7].

The rapid advancement of artificial intelligence, particularly deep learning, offers promising potential in neuroimaging and diagnosis of brain diseases. Data fusion techniques in deep learning have been employed to explore the correlations between structural and functional brain imaging [8]. However, most research in this area has primarily focused on using multimodal information for predicting behavioral outcomes [9] or diagnosing brain disorders [10]. These studies highlight the increased accuracy in detecting brain disorders using multimodal techniques in neuroscience, demonstrating that information from different imaging modalities can be effectively integrated. However, the specific processes by which structural and functional modalities correlate, particularly in the context of specific brain disorders, remain poorly understood. This knowledge gap hinders a better understanding of how structural and functional aspects of the brain influence each other. Developing methods to identify common information shared between structural and functional imaging in brain disorders is critical, yet challenging, as it involves unraveling complex interactions that are not fully explained by current technologies. Nevertheless, existing evidence suggests a significant correlation between inter-subject variations in brain structural networks and those observed in resting-state fMRI networks [11]. This indicates an underexplored yet potentially fruitful area of research in understanding the complex interplay between the brain's structure and function.

Generative deep learning models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), have shown significant success in neuroimaging applications. They excel at simulating neuroimaging data, detecting disease-specific patterns, and enabling modality translation, such as converting T1-weighted to T2-weighted MRI scans [45], [46]. This proficiency suggests the existence of a potential unifying biological or structural principle in brain imaging, bridging the gap between different imaging techniques [12], [13], [14], [15], [16]. The differing techniques of these imaging modalities might intersect at a common point of neural information within the brain, which argues for a unified strategy to gain insights into brain function. However, the integration and synthesis of GM and fMRI data, particularly in deep learning applications, remain comparatively underexplored [17], [18]. This contrasts with the substantial advancements in GAN models that predominantly focus on medical image synthesis between different modalities [47], [48]. fMRI data is typically high-dimensional, posing greater complexity in data fusion and generation. As a result, converting high-dimensional fMRI into an FNC matrix using ICA algorithms for translation between structural and functional images has not been previously explored and is the focus of our investigation.

The emergence of the vision transformer (ViT) marks a significant shift in computer vision, particularly within the realm of neuroimaging [19]. This innovative model diverges from

traditional convolutional neural networks (CNNs) by incorporating mechanisms initially crafted for language processing, such as the self-attention mechanism, enabling a more nuanced and holistic image analysis [20]. ViT's architecture, which dissects images into patches for processing through multiple transformer layers, allows for an in-depth analysis, independent of an image segment's spatial location. However, its computational demands, particularly its $O(N^2)$ complexity, pose challenges, propelling the quest for more streamlined architectures [21].

As researchers seek to enhance the efficiency of ViT through methods like architecture pruning and knowledge distillation without compromising performance [22], [23], the value of ViT in neuroimaging becomes increasingly apparent. The need for precision in brain scans calls for advanced, yet efficient, models, making ViT an attractive option due to its superior interpretability facilitated by attention mechanisms. Notably, ViT has been shown to outperform conventional CNNs in handling complex medical imaging datasets, which raises the possibility that ViT could replace CNNs as the foundational architecture in GANs. Such a shift could utilize ViT's detailed representational capabilities, which could be particularly advantageous for complex neuroscientific studies like attention mapping. By capitalizing on ViT's precision and computational efficiency, it could pave the way for significant breakthroughs in detecting and visualizing specific biomarkers within brain regions, thereby enhancing our understanding of the links between structural and functional neuroimaging. However, for high-dimensional 3D GM images, how to reduce the size of patches to obtain more detailed and accurate attention maps without significantly increasing the computational complexity of the ViT self-attention module is an important issue addressed in this paper.

To solve the problem of structural to functional brain image synthesis, this paper describes 1) the creation of a new conditional GAN model, which is called cEViT-GAN, that can generate functional connectivity matrices from GM data. As the generator and discriminator, an efficient ViT model is used. Since we utilize the attention maps generated by the ViT to identify potential biomarkers of schizophrenia in the brain's structure, we set smaller patches. However, our model remains capable of efficient training and operation. 2) In contrast to conventional self-attention operations, we select a block-wise self-attention layer that significantly reduces the computational cost without compromising performance. The mechanism for block-wise self-attention is versatile. The model accentuates regional relationships by employing self-attention operations in each block, thereby capturing localized patterns and interactions. In contrast, when inter-block self-attention is enabled, it ensures that long-term dependencies across the larger structure are not neglected. This dual strategy is ideal for identifying brain biomarkers from GM data. Focusing on specific regional relationships that indicate certain conditions or abnormalities is essential, but it is also necessary to analyze the entire brain image to completely comprehend and diagnose the issue. 3) To enhance training efficiency, we employ a pretrained ViT patch embedding layer, which was derived from an upstream task of diagnosing schizophrenia using the same model architecture. This allows the GAN model's generator to effectively extract already learned features, thereby improving the training efficiency. 4) In addition, our GAN model has the potential to be used as a biomarker identification tool for identifying the structural and functional connections of the human brain, particularly for various brain diseases such as schizophrenia.

II. RELATED WORKS

Generative adversarial networks (GANs), initially proposed by [24] and extended by [25], have significantly advanced as essential AI tools in various generative tasks. These tasks include

image and signal generation [26], as well as text-to-image and image-to-image synthesis [27]. In medical imaging, GANs play a crucial role in super-resolution, where they enhance image clarity and detail, and in the generation of synthetic images. The creation of these synthetic images is fundamental for data augmentation, training simulations, and the provision of enhanced diagnostic insights without the need for additional radiation exposure or patient involvement [28] [15]. Traditionally, GANs have predominantly utilized CNNs for both the generator and discriminator. However, the emergence of ViTs has led researchers to investigate more efficient architectures for ViT-based GAN models [29] [30]. ViTs are particularly effective in brain imaging, excelling at capturing comprehensive brain patterns, thus ensuring a more complete representation and superior feature extraction [49], [50]. This capability is especially beneficial in recognizing complex neural structures, surpassing the performance of CNNs. For instance, [31] introduced a pre-trained ViT model for classifying brain tumors, addressing the limitations of CNNs that tend to focus predominantly on minute pixel variations. Additionally, [32] demonstrated an enhanced ViT architecture capable of utilizing both structural and functional MRI data for predicting various stages of Alzheimer's disease. Furthermore, the integration of ViT and GAN has emerged as a novel trend in medical imaging. An example of this is the study by Zhao et al. [33], who developed a swin transformer-based GAN model [34] aimed at effective reconstruction of high-resolution MRI images.

In the domain of medical image synthesis, the focus has been on generating images across different modalities, such as CT, MRI, PET, and others. Dalmaz et al. [35] created a new GAN model that combines CNNs with transformer blocks. This model makes it much easier to make medical images that are similar and work better. However, we were hardly able to find related works that corresponded to MRI structural and functional image synthesis, besides our previous works, which synthesized FNC data from given sMRI and achieved a high correlation between real FNC and generated FNC data [36]. However, our previous use of a basic ViT-based GAN architecture was time-consuming and did not include the generation of structural biomarkers, which was a significant shortcoming.

III. METHODS

Our methodology is an innovative combination of deep learning architectures that address the complex problem of generating functional neural connectivity (FNC) maps from 3D GM data. Our model cEViT-GAN was designed to learn and generate high-fidelity FNC representations. We propose an efficient block-wise self-attention technique to avoid the significant computational overhead typically associated with ViT's processing of small image patches. This personalized strategy preserves ViT's tremendous feature extraction capabilities while maintaining computational efficiency, allowing the model to handle the large amounts of data associated with GM. We enhance our methods by superimposing the attentional weights from each layer of the ViT encoders onto the spatial information of 3D GM images, thereby aiding in the creation of sophisticated attention maps that not only reflect activations but also differentiate brain patterns between schizophrenia (SZ) and healthy control (HC) participants. Our technique paves the way for more insightful neuroimaging studies, potentially aiding early diagnosis and intervention efforts for mental health problems, by providing a visual and quantitative differentiation between groups.

A. Generative Adversarial Networks

Integrating generative adversarial networks (GANs) into the domain of medical imaging

necessitates a nuanced understanding of their loss functions. For our specific application of synthesizing FNC maps from 3D GM data, we construct a composite loss function that ensures the generation of realistic and medically informative images. The total loss $\mathcal{L}_{total}$ of our GAN framework is a weighted sum of four components:

$$\mathcal{L}_{total} = \mathcal{L}_G + \mathcal{L}_D + \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{corr}, \quad (1)$$

where $\mathcal{L}_G$ denotes the generator loss, $\mathcal{L}_D$ the discriminator loss, $\mathcal{L}_{MSE}$ the mean squared error loss, and $\mathcal{L}_{corr}$ the correlation loss. The terms λ_1 and λ_2 are hyperparameters that balance the contribution of the MSE loss and the correlation loss, respectively.

The generator loss $\mathcal{L}_G$ is defined as:

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log D(G(\mathbf{z}))], \quad (2)$$

where G is the generator, D is the discriminator, and $\mathbf{z}$ is a point sampled from the generator's input noise distribution $p_{\mathbf{z}}(\mathbf{z})$. The discriminator loss $\mathcal{L}_D$ is formulated as:

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))], \quad (3)$$

where $\mathbf{x}$ represents real data samples from the distribution $p_{data}(\mathbf{x})$. The mean squared error loss $\mathcal{L}_{MSE}$ is incorporated to penalize the pixel-wise differences between the generated and real images, thus preserving the structural integrity of the FNC maps:

$$\mathcal{L}_{MSE} = \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\| G(\mathbf{z}) - \mathbf{x}\|^2\right]. \quad (4)$$

The innovation in our approach is embodied by the correlation loss $\mathcal{L}_{corr}$, which ensures that the statistical dependencies between regions in the generated FNC maps are reflective of the true data. This is crucial for maintaining the biological fidelity of the neural connectivity patterns. The correlation loss is defined as:

$$\mathcal{L}_{corr} = 1 - \frac{\mathbb{E}\left[\left(\mathbf{x} - \mu_{\mathbf{x}}\right)\cdot\left(G(\mathbf{z}) - \mu_{G(\mathbf{z})}\right)\right]}{\sigma_{\mathbf{x}}\sigma_{G(\mathbf{z})}}, \quad (5)$$

where μ and σ denote the mean and standard deviation, respectively. This loss encourages the generated maps to have a correlation structure similar to that of the real FNC maps.

Our GAN architecture also incorporates a conditional input, whereby the generator receives both a sample of noise z and a label indicating the class (SZ or HC). This guides the generator towards producing FNC maps that are not only realistic but also correctly aligned with the specified condition:

$$G(\mathbf{z}|y) \quad \text{where} \quad y \in \{\text{SZ}, \text{HC}\}. \quad (6)$$

In essence, by carefully crafting the loss function and incorporating conditionality, our method aims to drive the

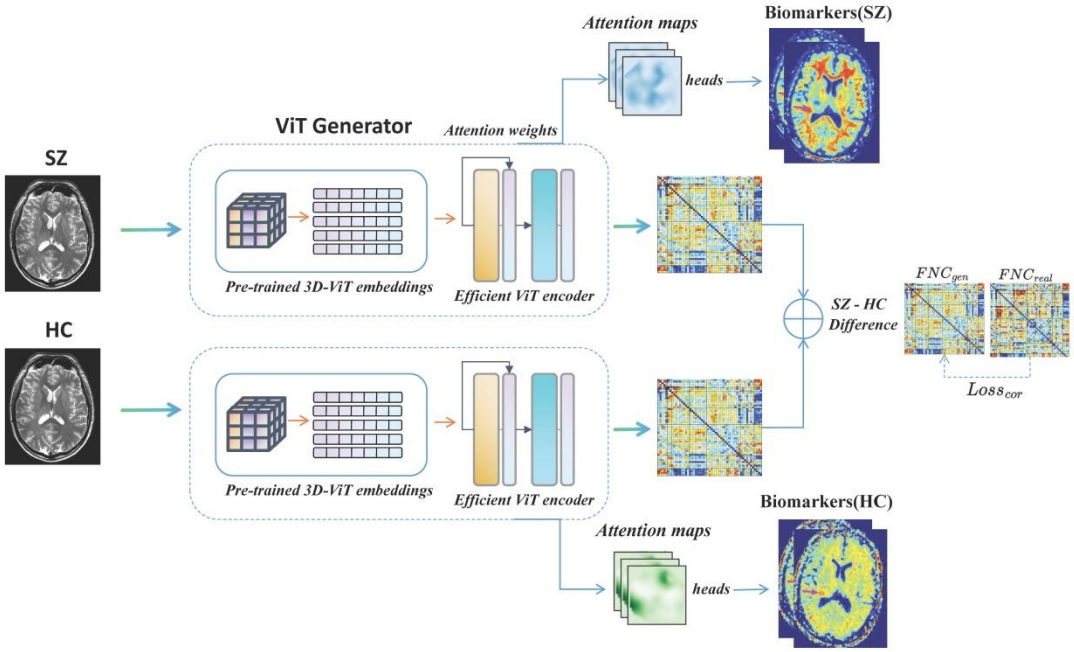

Fig. 1. The proposed methodology involves the analysis of brain MRI scans, specifically those labeled as SZ and HC. The objective is to generate group difference FNC data by utilizing an well-trained efficient generator from the cEViT-GAN framework. Additionally, attention weights are extracted from the ViT encoder to obtain 3D GM attention maps for the different groups. We then apply this approach to identify biomarkers associated with schizophrenia.

GAN towards producing medically valuable outputs.

B. Vision Transformer

The vision transformer (ViT) innovatively adapts transformer mechanisms, originally designed for natural language processing, to computer vision by treating image patches as a sequence of tokens and applying self-attention mechanisms to capture global dependencies within the image.

1) Pre-trained 3D Patch Embedding: The utility of pretrained models in deep learning is unparalleled, particularly in domains where data is scarce or where training from scratch is computationally prohibitive. Leveraging a pre-trained 3D ViT model, our generator benefits from an advanced starting point. This model, initially trained on upstream tasks such as the classification of SZ and HC from GM, has already learned a rich hierarchy of features that are highly relevant to our target domain. The pre-trained model forms the cornerstone of our generator's architecture. Specifically, for the patch embedding process, we utilize the pre-trained embeddings, denoted as:

$$\mathbf{E} = \mathrm{Pre}\left(\left[\mathbf{P}_1; \mathbf{P}_2; \ldots; \mathbf{P}_N\right]\right), \qquad (10)$$

where $\mathbf{P}_i$ corresponds to the flattened vector of the *i*-th 3D patch, and N is the number of non-overlapping 3D patches extracted from the GM input. The function Pre (·) encapsulates the process of obtaining the embedded representations using the pre-trained ViT model.

These pre-trained patch embeddings already encode the spatial hierarchies learned from the upstream classification task, providing a richly structured feature space that is finetuned for the generator:

$$\mathbf{E}^{*} = \mathbf{E} + \mathbf{E}_{pos}, \qquad (11)$$

where $\mathbf{E}^{*}$ represents the embeddings that will be utilized in the transformer encoder, and $\mathbf{E}_{pos}$ is the positional encoding added to the pre-trained embeddings.

These embeddings serve as the input to the ViT encoder, which comprises multiple layers of multi-headed self-attention and feed-forward networks:

$$\mathbf{T} = \mathrm{ViTEncoder}(\mathbf{E}), \qquad (8)$$

where **T** denotes the sequence of transformer encoder outputs corresponding to each patch embedding. Subsequently, each token produced by the ViT encoder is passed through a multilayer perceptron (MLP) network. This MLP is designed to reconstruct the small patches of the generated FNC matrix, transforming the abstract representations learned by the ViT into spatially structured outputs:

$$\mathbf{FNC}_{pa_i} = \mathrm{MLP}\left(\mathbf{T}_i\right). \qquad (9)$$

The collection of FNC patches $\mathbf{FNC}_{pa_i}$ is then reassembled to form the complete FNC map, which serves as the generator's final output:

$$\mathbf{FNC}_{gen} = \mathrm{Reassemble}\left(\left\{\mathbf{FNC}_{pa_1}, \mathbf{FNC}_{pa_2}, \ldots, \mathbf{FNC}_{pa_N}\right\}\right).$$

For the discriminator, the 3D ViT discerns between the real and generated FNC maps, employing a similar patch-based approach to extract features and perform classification. The discriminator's role is to evaluate the authenticity and quality of the generated FNC maps, guiding the generator through the adversarial training process to produce outputs that are increasingly indistinguishable from the real FNC maps derived from GM data. By integrating the ViT model into both the generator and discriminator of our GAN, we harness its potent capacity for capturing intricate patterns and dependencies within the complex data structure of three-dimensional brain imaging.

2) Block-wised Multi-head Self-attention: Incorporating the block-wise multi-head self-attention (BMHSA) [37] mechanism into our model optimizes computing efficiency while keeping the delicate features required for high-resolution biomarker detection from 3D GM data. We used BMHSA in vision tasks because of its excellent performance in dealing with long-text in NLP tasks. BMHSA partitions the collection of 3D GM patch embeddings into smaller, computationally efficient chunks, facilitating focused self-attention within these subdivisions to handle the small patch sizes essential for maintaining resolution in biomarker analysis of 3D GM data. Within each block, BMHSA operates by computing self-attention independently, which drastically reduces the overall computational load compared to traditional methods. Mathematically, the self-attention within a block *b* can be expressed as:

$$\text{Attn}\left(Q_b, K_b, V_b\right) = \text{softmax}\left(\frac{Q_b K_b^T}{\sqrt{d_k}}\right) V_b, \qquad (13)$$

where Q_b, K_b, and V_b are the queries, keys, and values for the block *b*, and d_k represents the scaling factor for the dot products within the softmax function to ensure numerical stability.

Leveraging the concept of multi-head attention, BMHSA allows the model to concurrently attend to different representational subspaces and positions within each block, formulated as:

$$\text{BMHSA}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O, \qquad (14)$$

$$\text{where head}_i = \text{Attn}\left(QW_i^Q, KW_i^K, VW_i^V\right), \qquad (15)$$

with each W_i^Q, W_i^K, and W_i^V denoting the respective parameter matrices for each attention head *i*, and W^O being the output linear transformation matrix.

The BMHSA approach ensures the emphasis of intra-block (regional) relationships while facilitating the preservation of inter-block (long-range) dependencies. These long-range dependencies are crucial for the analysis of structural brain images, as they allow the model to piece together localized information to form a comprehensive understanding of the brain's structure:

$$\mathbf{T} = \text{Concat}\left(\text{BMHSA}\left(\mathbf{E}_1\right), \ldots, \text{BMHSA}\left(\mathbf{E}_B\right)\right) + \mathbf{E}_{pos}, \qquad (16)$$

In this equation, **T** is the output of all the transformer encoder layers put together. It includes both detailed and general information about the brain's structure. The $\mathbf{E}_B$ terms show the embeddings from each block, and the $\mathbf{E}_{pos}$ terms show the positional encodings that are needed to keep the 3D MRI data's natural spatial relationships.

BMHSA Complexity Analysis: By employing the block-wise multi-head self-attention (BMHSA) mechanism, our model achieves significant reductions in computational costs while successfully generating high-resolution attention maps. Traditional self-attention mechanisms, such as those used in ViT models, exhibit a computational complexity that scales quadratically with the length of the sequence *n*. This complexity is expressed as $O\left(n^2 \cdot d\right)$, where *d* is the dimensionality of the attention heads. For long sequences, this scaling becomes computationally prohibitive.

BMHSA addresses this issue by partitioning the input sequence into smaller, fixed-size blocks, each of length *k*. Within each block, self-attention is computed independently, leading to a complexity of $O\left(k^2 \cdot d\right)$ per block. If the input sequence is divided into *m* such blocks, with the total sequence length *n* being equal to $m \times k$, the initial thought would be to express the overall complexity as the sum across all blocks, leading to $O\left(m \cdot k^2 \cdot d\right)$.

However, a more accurate representation of BMHSA's complexity takes into account the parallelizability of these block computations. Since each block's computation is independent, the per-block complexity of $O\left(k^2 \cdot d\right)$ remains, but the computations across different blocks can be performed in parallel. Therefore, the overall computational load does not directly scale with the number of blocks *m*.

Thus, the total computational complexity of BMHSA can be more accurately described as:

$$O\left(k^2 \cdot d\right) \times \text{ parallelization factor,}$$

In conclusion, by judiciously choosing an appropriate block size *k*, BMHSA effectively balances the trade-off between manageable computational costs and the granularity of attention required for detailed analysis in tasks such as high resolution biomarker detection from 3D GM data.

C. cEViT-GAN Architecture

The cEViT-GAN architecture, uniquely designed for analyzing 3D GM data and synthesizing FNC maps, stands out in the field of medical image processing by employing a purely self-attention mechanism instead of standard convolutional techniques. This purely ViT-based approach, in contrast to traditional CNN-based GAN architectures, as detailed in Table 1 which outlines the various layers and functions of our cEViT-GAN model. Figure 2 depicts the pipeline and overall architecture of cEViT-GAN.

Generator Architecture: The generator begins by taking small 3D GM patches, labeled as either SZ or HC . These patches are initially processed through pre-trained 3D embedding layers, utilizing the pre-trained ViT model to capitalize on its extensive feature extraction capabilities from GM data. The data then passes through BMHSA layers, which are crucial for efficient feature extraction and computational load management. The final stage involves MLPs reconstructing the FNC maps from these features, converting transformer outputs into spatially structured FNC patches, which are then assembled into a complete FNC map.

Discriminator Architecture: The discriminator's design features a pure 2D ViT that starts by segmenting FNC maps into patches and processing them through the ViT encoder, effectively discerning patterns to classify the input and produce a probability score indicating the authenticity of the FNC map, a crucial feedback mechanism for the adversarial training of the generator to create accurate and realistic FNC maps.

IV. EXPERIMENT

This section will describe the process of experimental setup, including the datasets and preprocessing, the training and testing of the models, the establishment of baselines, the implementation of the cEViT-GANs, and the experimental design to assess the structural and functional aspects of the brain.

A. Experimental Setups

1) Datasets: In our study, we utilized two comprehensive datasets pertinent to clinical schizophrenia research. Dataset 1 amalgamated data from three distinct studies: fBIRN (Functional Imaging Biomedical Informatics Research Network) across seven sites, MPRC (Maryland Psychiatric Research Center) spanning three sites, and COBRE (Center for Biomedical Research Excellence) at a single site. This aggregation culminated in a total of 827 participants, comprising 477 control subjects (average age: 38.76 ± 13.39, encompassing 213 females and 264

males) and 350 individuals diagnosed with schizophrenia (average age: 38.70 ± 13.14, including 96 females and 254 males). The fBIRN dataset was acquired using uniform resting-state fMRI (rsfMRI) parameters across all sites. We used a standard gradient echo-planar imaging (EPI) sequence with a repetition time (TR) of 2000 ms and an echo time (TE) of 30 ms. The voxels were 3.4375 × 3.4375 × 4 mm in size, and the field of view (FOV) was 220 × 220 mm. The data was captured using six Siemens Tim Trio 3-Tesla scanners and one General Electric Discovery MR750 3.0 Tesla scanner. In the COBRE segment, rsfMRI images were also taken using a standard EPI sequence, but with a slightly different TR/TE of 2000/29 ms and voxel sizes of 3.75 × 3.75 × 4.5 mm, within a field of view (FOV) of 240 × 240 mm, using a 3-Tesla Siemens Tim Trio scanner. The MPRC dataset was gathered using a trio of distinct 3-Tesla Siemens scanners, namely the Siemens Allegra, Trio, and Tim Trio.

Dataset 2 contained a total of 815 subjects, collected from several Chinese hospitals, including 326 subjects (age: 29.81 ± 8.68, females: 167, males: 159) of typical controls and 489 SZ individuals (age: 28.98 ± 7.63, females: 229, males: 260). The subjects were Chinese ethnic Han groups. The dataset was recruited from seven sites in China with the same recruitment criterion, including Peking University Sixth Hospital; Beijing Huilongguan Hospital; Xinxiang Hospital Simens; Xinxiang HospitalGE; Xijing Hospital; Renmin Hospital of Wuhan University; Zhumadian Psychiatric Hospital [51]. The resting-state fMRI data were collected with the following three different types of scanners across the seven sites: 3.0 Tesla Siemens Tim Trio Scanner, 3.0 T Siemens Verio Scanner, and 3.0 T Signa HDx GE Scanner (TR/TE = 2000/30 ms, voxel spacing size = 3 × 3 × 3 mm, FOV = 220 × 220 mm, and 480/360 volumes). Subjects were instructed to relax and lie still in the scanner while remaining calm and awake.

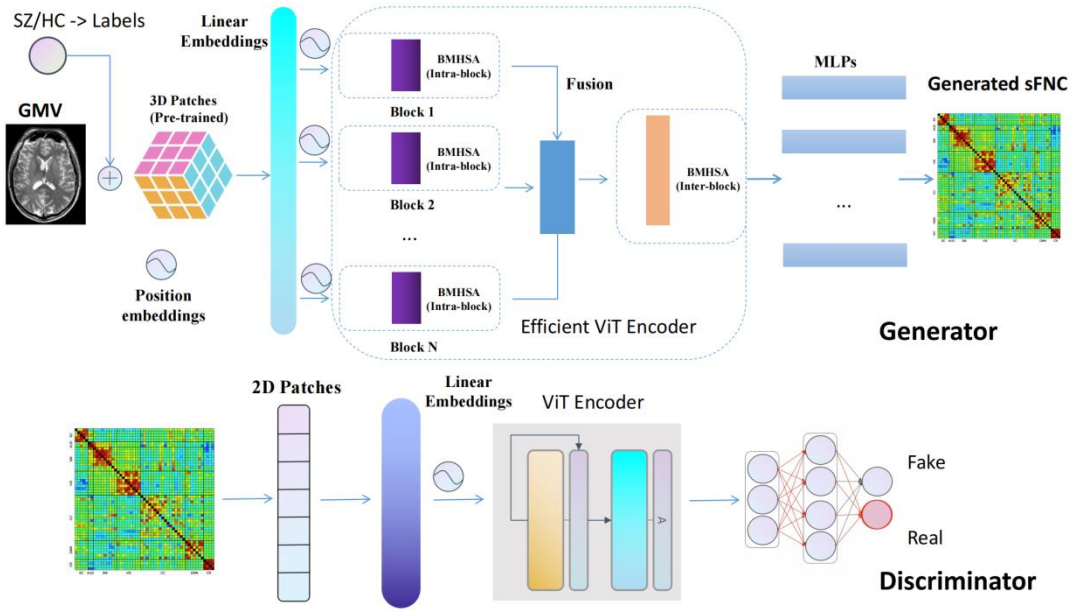

Fig. 2. cEViT-GAN's detailed architecture: The 3D GM and its label (SZ/HC) are input into the cEViT-GAN generator, passing through pre-trained 3D embedding layers and an efficient ViT encoder, followed by MLP outputs to form the FNC. The discriminator architecture is similar to a typical 2D ViT.

2) Pre-processing: To prepare the fMRI data, several critical processes were required: slice

timing correction, realignment, normalization to the EPI template, and smoothing with a 6 mm kernel. Our prior studies contain detailed descriptions of these preprocessing methods. Furthermore, FNC data was obtained using fMRI time series cross-correlation analysis. As spatial priors, a fully automated spatially limited ICA method and the NeuroMark template [38] were utilized. We used a voxel-based morphometry process on the sMRI data to acquire voxel-level GM data.

B. Models

1)Baselines: The primary goal of our comprehensive investigation of the efficacy and performance of several GAN models was to evaluate these models in terms of image-generating capabilities and output quality. The baseline models for comparison were carefully chosen, with special consideration given to their relevance to our pioneering work in synthesizing FNC from GM. While there are no clear previous works for synthesizing FNC, the closest similarity is found in the realm of image synthesis. As a result, we chose GAN models known for their expertise in this field as our baselines.

The first group of baselines includes CNN-based GAN models like Pix2Pix and deep convolutional GAN (DCGAN). The Pix2Pix model, which employs a U-Net generator and a PatchGAN discriminator, is well-known for its ability to solve image-to-image translation problems. The importance of this model in our research arises from its demonstrated ability to generate high-fidelity images from input images, a process that is like our goal of FNC synthesis from GM data. The integration of low and high-level characteristics in the generator by the U-Net architecture improves the detail and quality of the output images. Furthermore, the PatchGAN discriminator focuses on judging the realism of local image patches, which contributes greatly to image sharpness and overall coherence. As a result, these models provide a solid foundation for assessing the potential of GANs in our groundbreaking effort to synthesize FNC from GM. Moreover, we use traditional self-attention-based cViT-GAN as other baselines, which can show the efficiency of our model.

2) cViT-GAN: We employ the traditional cViT-GAN as another baseline for ablation studies, which utilizes conventional self-attention mechanisms. This comparison aims to demonstrate the distinct lightweight advantages of our ViT-encoder that incorporates core blockwise multi-head self-attention (BMHSA) layers. This experiment not only underscores our design's ability to reduce training time and computational complexity but also confirms that our lightweight approach maintains training accuracy despite the simplifications.

3)cEViT-GANs: Our research into novel GAN models resulted in the creation of the cEViT-GAN framework, a revolutionary technique developed exclusively for FNC synthesis. The cEViT-GAN models incorporate cutting-edge approaches, including a pre-training strategy focused on embedding 3D

TABLE I
CEVIT-GAN ARCHITECTURE OVERVIEW

Component	Layer/Function	Description
Generator	Input	Processes 3D GM patches labeled SZ or HC
	Pre-trained 3D Embedding	Utilizes pre-trained ViT model for feature extraction
	BMHSA Layers	Manages computational load, extracts features
	MLPs for Reconstruction	Converts encoder outputs to structured FNC

		patches
	Output Assembly	Assembles patches into complete FNC map
Discriminator	Patch Embedding	Divides FNC maps into patches for processing
	ViT Encoder	Processes embedded patches, extracts features
	Classification Output	MLP that outputs probability of real or generated map

patches and efficient usage of ViT blocks via blockwise self-attention. This novel combination intends to improve picture synthesis quality by combining the strengths of CNNs and transformers.

To optimize cEViT-GAN for our specific needs, we introduced several modifications. The first, cEViT-GAN-b3, consists of three parallel BMHSA blocks without interblock self-attention, prioritizing speed and efficiency while still delivering high quality images. In addition, the cEViT-GAN-b3large includes an interblock self-attention mechanism to enhance the model's ability to capture and integrate complex data patterns for more accurate FNC synthesis. In addition, the cEViT-GAN-b6 uses six parallel BMHSA blocks without inter-block connections to explore the effects of increased parallelism on computational efficiency and image quality. Our most advanced configuration, the cEViT-GAN-b6large, forms the cornerstone of our study and serves as the basis for all visualization and analysis. This model combines multiple BMHSA layers with interblock self-attention, designed to balance precise feature acquisition with efficient processing of 3D GM data. The inclusion of interblock self-attention is critical, as it allows for more effective integration of information across layers, which can lead to more refined and accurate synthesis of FNC from GM data.

E. Experiments Details

1) Pre-training: When testing our ViT-based GAN models, including the baseline and our cViT-GAN variants, we utilize pre-trained 3D linear embeddings. These embeddings are obtained from a previously developed multimodal deep learning model designed for schizophrenia diagnosis and classification [17]. The significance of this pre-training is particularly pronounced for the generator components of our GANs, facilitating an efficient transfer of learned features across medical imaging tasks. This enhances the generalizability and robustness of our models.

The generators in our cEViT-GANs benefit significantly from starting with weights derived from these pre-trained embeddings. This not only accelerates the training process by providing an informed initialization but also improves the models' overall efficiency and effectiveness. The embeddings encapsulate a wealth of features relevant to schizophrenia, enriching the generators with nuanced neuroimaging patterns associated with the condition. Consequently, our GAN models can synthesize FNC images from GM data that are more detailed, and clinically relevant to our neurological focus.

2) Train and Validation: It's worth noting that while all CNN-based GAN models employ a uniform set of parameters and training techniques, ViT-based GANs, including baseline and cEViT-GAN variations, utilize a distinct set. For CNN-based GANs, we use Kaiming initialization for selecting initial weights and the AdamW optimizer for both the generator and the discriminator, setting the learning rate at 1e-3 with a MultistepLR schedule that adjusts at the 20th, 50th, and 150th epochs. Conversely, ViT-based GAN models, which incorporate pre-trained weights for 3D patch embedding in the generator, require a lower learning rate of 1e-4, also with AdamW as the

optimizer and a MultistepLR schedule making adjustments at the 20th, 50th, and 90th epochs. All models are trained with a batch size of 32, and the pre-training stage typically leads to convergence around the 90th epoch.

Using cross-validation on both types of models improves model resilience and dependability. This technique is useful for determining how the models would perform on different data sets, decreasing the danger of over-fitting and assuring generalization. Our training and validation operations are powered by 8 NVIDIA Tesla V100 GPUs, and we utilize PyTorch as our model framework. We adopt parallel and distributed training approaches, specifically using PyTorch's distribution methods, to distribute the training load across multiple GPUs. This strategy not only enhances processing efficiency but also significantly reduces training times. It is particularly beneficial for handling the large volumes of data and complex neural network architectures required in our research. In addition, we use PyTorch's built-in distribution strategies, which use the all-reduce algorithm rather than a parameter server approach. This method efficiently aggregates gradients across multiple GPUs to ensure synchronized updates and optimal training performance.

F. Evaluation Metrics

In our research, we deploy a trio of critical metrics to assess the efficacy of our model in synthesizing FNC patterns. These metrics include the Mean Squared Error (MSE), the Pearson Correlation Coefficient, and the Cosine Similarity. Each of these metrics plays a crucial role in evaluating the precision and reliability of the FNC patterns generated by our model, offering distinct insights into the model's performance and facilitating a comprehensive assessment when compared to authentic FNC data.

1) Mean Squared Error (MSE): MSE is a metric commonly utilized in regression analysis and signal processing. It quantifies the average of the squares of errors, which are the differences between the estimated values and the actual values. In the context of our FNC data, the MSE is calculated as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Here, n denotes the total number of FNC entries, Y_i represents the actual FNC value, and $\hat{Y}_i$ signifies the estimated FNC value produced by the model. A lower MSE value is indicative of superior model performance, signifying a reduced deviation from the true FNC values.

2) Pearson Correlation Coefficient (PCC): The PCC is a measure that quantifies the linear correlation between two datasets. It yields a value within the range of -1 to 1, where 1 denotes a total positive linear correlation, 0 signifies no linear correlation, and -1 indicates a total negative linear correlation. In evaluating our model, this coefficient is defined as:

$$r = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(\hat{Y}_i - \bar{\hat{Y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}}$$

Where $\bar{Y}$ and $\bar{\hat{Y}}$ are the mean values of the actual and estimated FNCs, respectively. A

higher absolute value of this coefficient implies a stronger correlation between the generated FNCs and the real data.

3) Cosine Similarity: Cosine Similarity is a metric employed to ascertain the similarity between two vectors, irrespective of their magnitude, and is especially pertinent in high-dimensional spaces. The cosine similarity between the actual and the model-generated FNC vectors is computed as:

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n} Y_i \times \hat{Y}_i}{\sqrt{\sum_{i=1}^{n} Y_i^2} \times \sqrt{\sum_{i=1}^{n} \hat{Y}_i^2}}$$

In this formula, the numerator represents the dot product of the actual and estimated FNC vectors, while the denominator is the product of the Euclidean norms of these vectors.

Together, these metrics provide a strong and flexible way to check how accurate and similar the FNC patterns our model creates are to real FNC data. This gives us important information about how well the model can copy complex neural connectivity patterns.

G. Visualizations

We performed intense brain anatomical and functional visualization based on self-attention operations in two phases. The first step is to use attention weights on original brain maps to find any biomarkers in GM data while also making a matching FNC. The second step is to compare the made FNC with the real FNC, which could help find functional biomarkers for SZ disease.

1) MRI Attention Maps: To extract attention weights, we used a rollout method, concatenating weights from each block from blockwise multihead self-attention and superimposing the weights of interblock self-attention onto the averaged block weights, producing a comprehensive attention map useful for detecting potential biomarkers in GM data. High attention weights help us identify regions of interest that may be connected with various neurological diseases, such as SZ. The attention map serves as a guide, showing the most important sections of the GM data. Researchers can get insights into the fundamental mechanisms of SZ and potentially other neurological illnesses by better understanding these areas. The incorporation of attention weights into GM data analysis represents a significant improvement in neuroimaging and the research of brain diseases. Fortunately, analyzing the group difference (HZ-HC) attention map allowed us to identify brain areas strongly associated with SZ that aligned with our existing knowledge. Following on from the analysis of group differences, our model also supports the generation of FNC maps based on individual cases of GM. This capability represents a promising avenue for the exploration of personalized biomarkers, as it allows the adaptation of our approach to individual variations in GM data. By tailoring the analysis to each individual subject, researchers can potentially uncover unique patterns and connections in brain structures that are specific to individual neurological profiles, increasing the precision and relevance of biomarker discovery in conditions such as SZ and other neurological diseases.

2) FNC Maps: Our cEViT-GANs are capable of generating reasonably accurate FNC maps. We generate FNC maps for each subject by conditioning on the GM data from each group (SZ and HC), and then average these maps to assess and validate the accuracy and effectiveness of our model. Further, by analyzing the averaged group differences in FNC maps, we demonstrate that our model can effectively learn these distinctions, using GM as structural input to derive functional differences, thereby substantiating the biological significance of the connection

between structure and function.

V. RESULTS

This section presents the outcomes of the experiments as well as a visualization of the structure and function of the brain using attention maps and FNC biomarkers. Initially, we conducted exhaustive experiments on various baselines and our cEViT-GAN variants to ensure that our model exhibited superior accuracy and robustness.

A. Model Performance

The experimental results shown in Figure 3 highlight the performance differences between our baseline models (Pix2Pix, DCGAN, and cViT-GAN) and our new cEViT-GAN variants. From these results, it is clear that the standard DCGAN, which uses a pure CNN backbone, underperforms in FNC generation, suggesting that pure CNN architectures are not particularly effective for this task. Pix2Pix, a well-known GAN model that adapts both the generator and the discriminator, meets the requirement for high quality image generation to some extent.

The use of a pure ViT backbone, as in cViT-GAN, is advantageous for the extraction of long-range features due to its self-attention mechanism, but results in higher computational costs. Further reduction of the patch size in this context could lead to memory overload. Our proposed cEViT-GAN, especially the cEViT-GAN-b6large model with interblock self-attention, shows excellent performance; however, the inclusion of interblock self-attention increases the training time. The base model of EViT-GAN, such as cEViT-GAN-b6, significantly reduces training complexity without compromising accuracy - maintaining the same level of accuracy as cViT-GAN but with reduced training time. Therefore, as we explore ways to further reduce patch sizes in the future, the use of EViT-GAN could not only reduce training times, but also maintain the quality and accuracy of the generated FNCs while improving the refinement of attention maps.

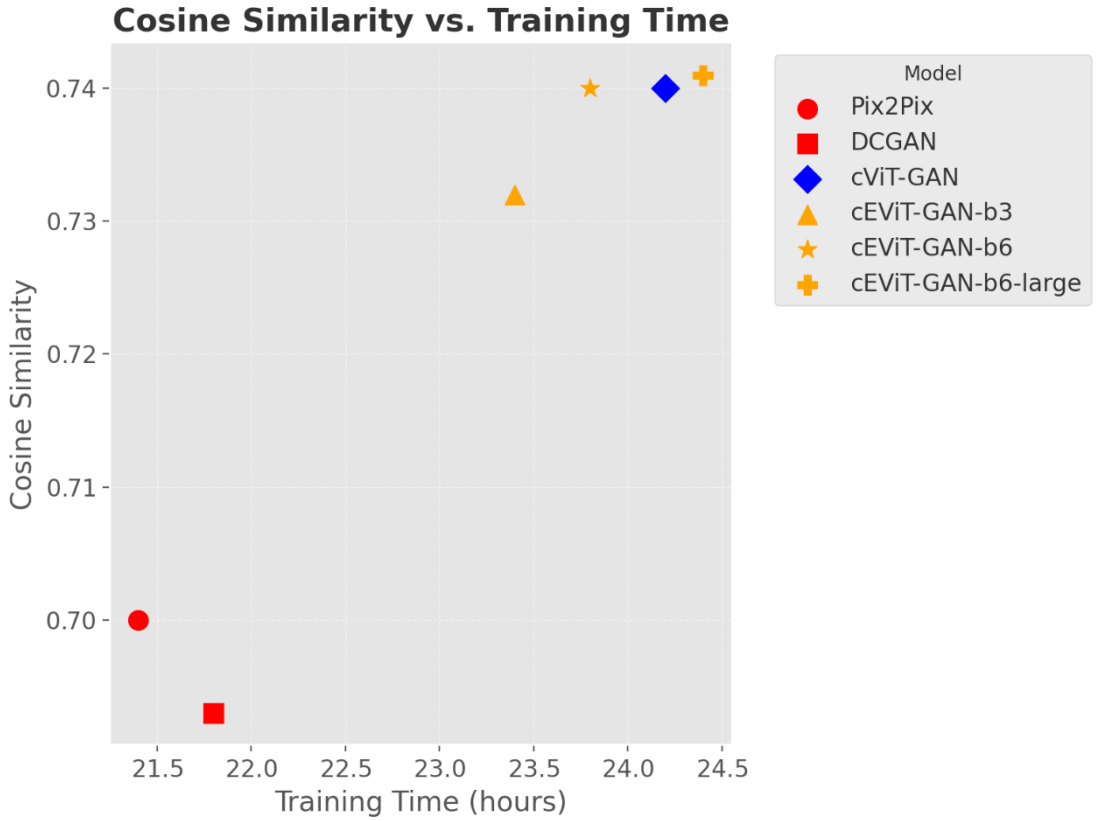

Fig. 3. Comparison of different models in terms of Cosine Similarity and Training Time.

TABLE II
MODEL PERFORMANCE COMPARISON

Model	Backbone	Cosine Similarity	Pearson Correlation	Training Time (hours)
Pix2Pix	CNN	0.7	0.71	21.4
DCGAN	CNN	0.693	0.693	21.8
cViT-GAN	ViT	0.74	0.74	24.2
cEViT-GAN-b3	ViT with BMHSA	0.732	0.731	23.4
cEViT-GAN-b6	ViT with BMHSA	0.74	0.741	23.8
cEViT-GAN-b6-large	ViT with BMHSA	0.741	0.741	24.4

B. MRI Attention Maps

We analyzed attention weights in our cEViT-GAN generator to generate our 3D GM attention maps. We created subject-specific attention maps for each member of our testing set, then tested for group differences using a two-sample t-test. Each voxel in our attention maps represents a t-value from this statistical test. To account for multiple comparisons, we used the false discovery rate (FDR) method with a $q < 0.05$ threshold. This approach accounts for the possibility of type I errors while running several statistical tests. The attention maps that arise emphasize areas with statistically significant changes in activation patterns between groups. Figure 4 shows the attention maps in a three-plane view.

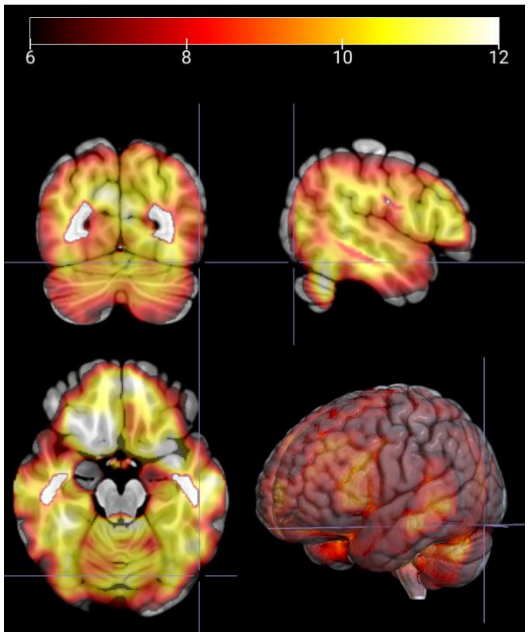

Fig. 4. The 3D MRI attention maps for group difference analysis (SZ vs HC), which indicate the significant ROIs that are strongly associated with schizophrenia disease.

Figure 4 indicates that while generating the related functional outputs, our model prioritized four brain regions: the medial pre-frontal cortex (mPFC), the dorsolateral prefrontal cortex (DL-PFC), the temporal lobe, and the cerebellum. Schizophrenia is a diverse, complex psychiatric

condition that frequently involves dysfunctions in numerous brain circuits. Based on traditional neuroscience and previous knowledge, mFPC is intimately related to executive processes and decision-making, both of which can be affected in schizophrenia [39]. The mPFC is also involved in emotional processing, and abnormalities here can be linked to negative schizophrenia symptoms including apathy and social disengagement [40]. DL-PFC is required for cognitive control and working memory, both of which are frequently impaired in people with schizophrenia. Deficits in this area can contribute to the disorder's hallmarks of disorganized thinking and trouble focusing attention. The superior temporal gyrus, in particular, is connected with auditory processing and language. Temporal lobe dysfunction has been linked to auditory hallucinations and language difficulties seen in schizophrenia patients [41]. Finally, the cerebellum's significance in cognitive processing is now recognized. Cerebellar abnormalities may contribute to cognitive impairments and affective dysregulation in schizophrenia, according to recent research [42], [43].

C. FNC Biomarkers

1) FNC Analysis: In this study, a sophisticated GAN model was employed to generate FNC outputs from a test dataset. Our analysis revealed that the model's output for the whole average FNC exhibited a strong correlation (0.97) with the actual FNC data across all subjects. This is effectively visualized in Figure 5, which compares the model-generated whole average FNC with the genuine FNC data. The ability of our GAN model to replicate FNC from 3D MRI scans of GM with high accuracy can be credited to the identification of neural structures via independent component analysis (ICA). ICA has been known to uncover network-like structures within the resting gray matter, providing insights into the model's capability to replicate these intricate neural patterns. The correlation observed in our model's output with the real data not only validates our approach but also aligns it with previous scientific research in neuroimaging [44], [16], [11].

2) Group Difference Analysis: We also show the produced and real group-difference FNC (HC-SZ). Figure 5 shows a comparison of calculated and actual FNC group differences. Our model can infer group-difference FNC from brain structure with a remarkably high correlation (0.74) especially given brain function contains unique information above and beyond brain structure. Our cEViT-GAN model can identify a strong similarity between the generated group-difference FNC and the real one, and the patterns are those that are know to be implicated in schizophrenia, including subcortical areas. These include connections between the cerebellum and the subcortical (CB-SC), auditory (CB-AUD), somatomotor (CB-SM), visual (CB-VS), cingulo-opercular (CB-CC), default mode (CB-DM), and the cerebellum itself (CB-CB). The synthetic FNC data obtained by GM has a remarkable correlation with real FNC data, with similarities reaching 0.85 in certain subcortical linkages.

This important finding shows that subcortical structures are important for identifying differences between HC and SZ participants and that the cEViT-GAN model has a good performance of showing these important structural-functional relationships. The remarkable similarity in subcortical areas shows that our model is quite good at reproducing complicated, potentially clinically relevant brain patterns. Such capabilities signal new opportunities to improve our understanding of diseases such as schizophrenia, to offer more precise diagnostic measures, and to personalize therapy methods. Furthermore, our model demonstrates that there is a high level of agreement in the difference in values for other pairs of connections, such as cingulo-opercular (CC-CC), somatomotor-default mode (SM-DM), and visual-default mode (VS-DM). We also find moderate parallelism in visual-auditory (VS-AUD) and

cingulo-opercular-somatomotor (CC-SM) pairs. These findings provide greater insight into how the disparities in FNC observed between the HC and SZ groups may be caused by underlying structural issues. The combined insights are critical for developing more refined diagnostic tools and therapy approaches for navigating the complexities of schizophrenia.

3) Cross-domain Analysis: Our FNC matrix cross-domain analysis provides a more detailed view of the link between structural and functional data. The produced and real FNC matrices have a total similarity measure of 0.74, which shows that there is a significant relationship, but the structural data does not fully reflect all functional features. The complicated nature of brain functionality, which cannot be entirely extrapolated from structural imaging, may account for this disparity.

Upon examining the cross-domain correlations, it becomes apparent that the within-domain correlations, such as AUD-AUD, exhibit a remarkably high similarity (0.955). This suggests that the structural data accurately reflects the functional connection of the auditory network. This is supported by strong correlations in subdomains like SC-AUD (0.847) and SC-SM (0.824), which show stable structural-functional alignment in the motor function and sensory processing domains.

The cEViT-GAN model captures the cerebellum's constant functional patterning, as evidenced by its strong intra-domain correlation (CB-CB at 0.821). Cross-domain interactions, like those between the default mode network and the cerebellum (DMN-CB) and the somatomotor and cerebellar regions (SM-CB), have moderate to high correlations. This means that the model can show how different parts of the brain work together. Notably, the lower correlations in coupling between other regions including SC-CB (0.160) and AUD-CB (0.053) show the challenge of mapping functional networks from structural data, especially when there are complicated connections between regions. These areas may indicate distinct functional characteristics or dynamic interconnections that are not readily apparent in GM data. These correlations are specific across domain sizes, from the small 2x2 matrices to the large 17x17 matrices. This makes them useful for checking the authenticity of FNC representations that have been made. Furthermore, it identifies areas where the generative model's performance could be enhanced to more accurately replicate the intricate tapestry of human brain connectivity.

Finally, our findings highlight the benefits and drawbacks of employing cEViT-GAN to replicate FNC matrices using structural data. The model's high fidelity in some domains encourages its use in clinical settings, whereas inequalities in others call for further research into the multidimensional nature of brain structure-function interactions.

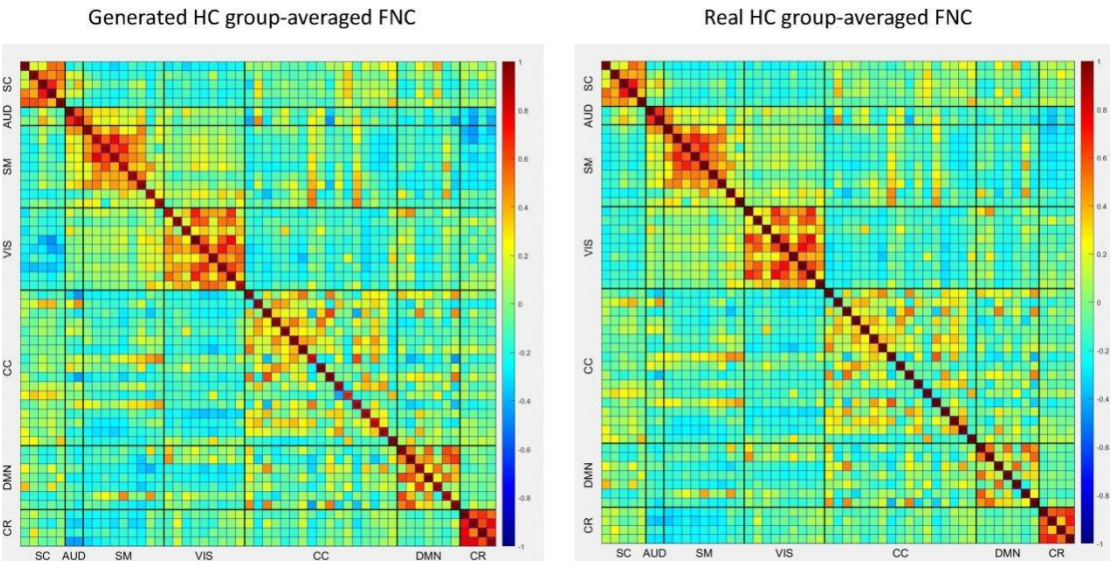

Fig. 5. The generated whole average FNC vs. real whole average FNC

D. Structural-to-functional Connectivity

The identification of biomarkers in SZ by combining structural and functional neuroimaging data tells a captivating story about the disorder's neuropathology. The findings of our investigation show a significant agreement between structural and functional indicators, highlighting the complicated connection between brain structure and function in SZ.

Significant structural sections include the medial frontal cortex (mPFC), dorsolateral prefrontal cortex (DL-PFC), and cerebellum. Similar functional regions show significant changes in connection patterns, particularly in the default mode network (DMN) and auditory and somatomotor activities. This correspondence between structural changes and functional connectivity disturbances allows for a more comprehensive understanding of SZ pathophysiology. For example, functional connectivity disruptions in the mPFC and DL-PFC, which are important for executive functioning and cognitive control, coincide with structural alterations in these areas, contributing to the cognitive and affective dysregulation seen in SZ patients. Both structural and functional findings highlight the importance of the cerebellum in SZ, an area that has been understudied until now. Changes in cerebellar areas correspond structurally with changes in functional connectivity within the cerebellum and its linkages to other brain networks. This shows that the cerebellum may play a role in the larger network dysfunctions that characterize SZ, going beyond its traditional concept of motor control.

Furthermore, the temporal lobe, a region involved in auditory processing, exhibits both structural and functional abnormalities, which correspond to clinical symptoms such as auditory hallucinations, which are common in SZ. This is supported by the significant correlation in functional networks, including the auditory cortex (AUD), which mirrors the anatomical findings. These similarities hint at a more integrated model of SZ in which structural anomalies are not isolated but have a considerable impact on the functional network dynamics. This model supports the idea that SZ is a disorder of "disconnected connection," with the symptoms being caused by the interaction of damage to the structure and problems with the way the network works.

In conclusion, the convergence of structural and functional biomarkers in our work have provided some new insights into our understanding of SZ. It demonstrates the interrelated nature of structural and functional network changes, providing a more comprehensive view of the

disorder's neurobiological roots. We hope our understanding can be further increase by an integrative approach like this, pontentially leading to the development of more effective diagnostic tools and targeted treatment options that are tailored to the personalized nature of SZ.

VI. DISCUSSION AND CONCLUSION

In this study, we introduced the cEViT-GAN, a novel approach that combines GAN with ViT and a new lightweight blockwise multihead self-attention technique. This model effectively generates FNC matrices from brain structural GM data, supporting the neuroscientific and biological perspective that there is a link between brain structure and function.

In particular, in neurological disorders such as schizophrenia, changes in brain function are often due to underlying changes in brain GM structure. By analyzing the results generated on a per-subject basis, our use of attention map technology has enabled the pinpointing of brain structures, such as the medial prefrontal cortex (mPFC), dorsolateral prefrontal cortex (DL-PFC), and cerebellum, that drive functional changes. Furthermore, our model effectively simulates characteristics and changes similar to those found in real FNCs, providing potential evidence that these functional changes originate from specific brain structures. This is particularly evident when comparing generated FNCs with real FNCs.

However, our model has limitations. For example, the conditional generative model, which typically operates under the supervision of a target generative object, is influenced by that target and attempts to replicate its statistical properties. Consequently, the FNC matrices generated by our model are supervised by actual FNC data and are not generated solely on the basis of structural GM data. To isolate the unique information derived from FNC structural data, it is necessary to eliminate the influence of unconditional generation from actual FNC data, a factor not addressed in our experiments.

Despite these limitations, our model represents a pioneering exploration of the use of data-driven 3D structural data to generate high quality FNCs. Our findings on a schizophrenia dataset provide guidance for future work. In the future, we aim to further extend and validate our model to develop a more generalized pipeline that is potentially applicable to a broader range of brain disorders and datasets.

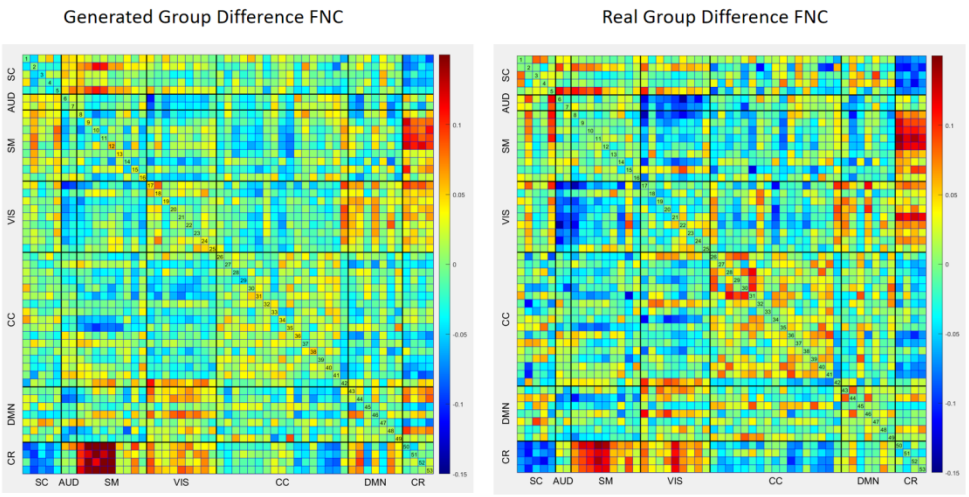

Fig. 6. The generated 'fake' group difference FNC matrix vs. real group difference FNC matrix.

Ethics for data: We are doing secondary analysis of data from public and private repositories. All data were collected under appropriate ethics approval and all subject signed informed consent.

CRediT authorship contribution statement: Yuda Bi: Investigation, Validation, Visualization, and Writing – original draft, Writing – review & editing. Anees Abrol: Formal analysis, Writing – review & editing. Sihan Jia: Writing – review & editing. Jing Sui: Writing – review & editing, Data curation. Vince D. Calhoun: Funding acquisition, Resources, Writing – review & editing.

REFERENCES

[1] G. D. Pearlson and V. Calhoun, "Structural and functional magnetic resonance imaging in psychiatric disorders," *The Canadian Journal of Psychiatry, vol*. 52, no. 3, pp. 158–166, 2007.

[2] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative *et al*., "Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database," *neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[3] U. Khatri and G.-R. Kwon, "Alzheimer's disease diagnosis and biomarker analysis using resting-state functional mri functional brain network with multi-measures features and hippocampal subfield and amygdala volume of structural mri," *Frontiers in aging neuroscience*, vol. 14, p. 818871, 2022.

[4] X. Zhao, C. K. E. Ang, U. R. Acharya, and K. H. Cheong, "Application of artificial intelligence techniques for the detection of alzheimer's disease using structural mri images," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 456–473, 2021.

[5] N. Franzmeier, N. Koutsouleris, T. Benzinger, A. Goate, C. M. Karch, A. M. Fagan, E. McDade, M. Duering, M. Dichgans, J. Levin *et al*., "Predicting sporadic alzheimer's disease progression via inherited alzheimer's disease-informed machine-learning," *Alzheimer's & Dementia*, vol. 16, no. 3, pp. 501–511, 2020.

[6] J. Oh, B.-L. Oh, K.-U. Lee, J.-H. Chae, and K. Yun, "Identifying schizophrenia using structural mri with a deep learning algorithm," *Frontiers in psychiatry*, vol. 11, p. 16, 2020.

[7] C. J. Honey, J.-P. Thivierge, and O. Sporns, "Can structure predict function in the human brain?" *Neuroimage*, vol. 52, no. 3, pp. 766– 776, 2010.

[8] V. D. Calhoun and J. Sui, "Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness," *Biological psychiatry: cognitive neuroscience and neuroimaging*, vol. 1, no. 3, pp. 230–244, 2016.

[9] J. Sui, R. Jiang, J. Bustillo, and V. Calhoun, "Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises," *Biological psychiatry*, vol. 88, no. 11, pp. 818–828, 2020.

[10] B. Rashid and V. Calhoun, "Towards a brain-based predictome of mental illness," *Human brain mapping*, vol. 41, no. 12, pp. 3468–3535, 2020.

[11] N. Luo, J. Sui, A. Abrol, J. Chen, J. A. Turner, E. Damaraju, Z. Fu, L. Fan, D. Lin, C. Zhuo *et al*., "Structural brain architectures match intrinsic functional networks and vary across domains: a study from 15 000+ individuals," *Cerebral Cortex*, vol. 30, no. 10, pp. 5460–5470, 2020.

[12] J. Pan, B. Lei, Y. Shen, Y. Liu, Z. Feng, and S. Wang, "Characterization multimodal connectivity of brain network by hypergraph gan for alzheimer's disease analysis," in *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4*. Springer, 2021, pp. 467–478.

[13] X. Dai, Y. Lei, Y. Fu, W. J. Curran, T. Liu, H. Mao, and X. Yang, "Multimodal mri synthesis using unified generative adversarial networks," *Medical physics*, vol. 47, no. 12, pp. 6343–6354, 2020.

[14] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "Gans for medical image synthesis: An empirical study," *Journal of Imaging*, vol. 9, no. 3, p. 69, 2023.

[15] Y. Liu, A. Chen, H. Shi, S. Huang, W. Zheng, Z. Liu, Q. Zhang, and X. Yang, "Ct synthesis from mri using multi-cycle gan for head-andneck radiation therapy," *Computerized medical imaging and graphics*, vol. 91, p. 101953, 2021.

[16] N. Luo, J. Sui, A. Abrol, D. Lin, J. Chen, V. M. Vergara, Z. Fu, Y. Du, E. Damaraju, Y. Xu *et al*., "Age-related structural and functional variations in 5,967 individuals across the adult lifespan," *Human brain mapping*, vol. 41, no. 7, pp. 1725–1737, 2020.

[17] Y. Bi, A. Abrol, Z. Fu, and V. Calhoun, "Multivit: Multimodal vision transformer for schizophrenia prediction using structural mri and functional network connectivity data," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.

[18] M. A. Azam, K. B. Khan, S. Salahuddin, E. Rehman, S. A. Khan, M. A. Khan, S. Kadry, and A. H. Gandomi, "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Computers in biology and medicine*, vol. 144, p. 105253, 2022.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al*., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking," *arXiv preprint arXiv:2309.02031*, 2023.

[22] Y. Tang, K. Han, Y. Wang, C. Xu, J. Guo, C. Xu, and D. Tao, "Patch slimming for efficient vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 165–12 174.

[23] X. Chen, Q. Cao, Y. Zhong, J. Zhang, S. Gao, and D. Tao, "Dearkd: data-efficient early knowledge distillation for vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 052–12 062.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[25] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401– 4410.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125– 1134.

[28] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. Á. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, “Madgan: Unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction,” *BMC bioinformatics*, vol. 22, no. 2, pp. 1–20, 2021.

[29] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, “Vitgan: Training gans with vision transformers,” *arXiv preprint arXiv:2107.04589*, 2021.

[30] S. Hirose, N. Wada, J. Katto, and H. Sun, “Vit-gan: Using vision transformer as discriminator with adaptive data augmentation,” in *2021 3rd International Conference on Computer Communication and the Internet (ICCCI)*. IEEE, 2021, pp. 185–189.

[31] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, “Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling,” *Current Oncology*, vol. 29, no. 10, pp. 7498– 7511, 2022.

[32] S. Sarraf, A. Sarraf, D. D. DeSouza, J. A. Anderson, M. Kabia, and A. D. N. Initiative, “Ovitad: Optimized vision transformer to predict various stages of alzheimer’s disease using resting-state fmri and structural mri data,” *Brain Sciences*, vol. 13, no. 2, p. 260, 2023.

[33] X. Zhao, T. Yang, B. Li, and X. Zhang, “Swingan: A dual-domain swin transformer-based generative adversarial network for mri reconstruction,” *Computers in Biology and Medicine*, vol. 153, p. 106513, 2023.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *in Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[35] O. Dalmaz, M. Yurt, and T. Çukur, “Resvit: Residual vision transformers for multimodal medical image synthesis,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[36] Y. Bi, A. Abrol, J. Sui, and V. Calhoun, “Cross-modal synthesis of structural mri and functional connectivity networks via conditional vitgans,” *arXiv preprint arXiv:2309.08160*, 2023.

[37] J. Qiu, H. Ma, O. Levy, S. W.-t. Yih, S. Wang, and J. Tang, “Blockwise self-attention for long document understanding,” *arXiv preprint arXiv:1911.02972*, 2019.

[38] Y. Du, Z. Fu, J. Sui, S. Gao, Y. Xing, D. Lin, M. Salman, A. Abrol, M. A. Rahaman, J. Chen *et al*., “Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders,” *NeuroImage: Clinical*, vol. 28, p. 102375, 2020.

[39] X. J. Chai, S. Whitfield-Gabrieli, A. K. Shinn, J. D. Gabrieli, A. Nieto Castañón, J. M. McCarthy, B. M. Cohen, and D. Öngür, “Abnormal medial prefrontal cortex resting-state connectivity in bipolar disorder and schizophrenia,” *Neuropsychopharmacology*, vol. 36, no. 10, pp. 2009– 2017, 2011.

[40] J. H. Callicott, A. Bertolino, V. S. Mattay, F. J. Langheim, J. Duyn, R. Coppola, T. E. Goldberg, and D. R. Weinberger, “Physiological dysfunction of the dorsolateral prefrontal cortex in schizophrenia revisited,” *Cerebral cortex*, vol. 10, no. 11, pp. 1078–1092, 2000.

[41] L. L. Davidson and R. W. Heinrichs, “Quantification of frontal and temporal lobe brain-imaging findings in schizophrenia: a meta-analysis,” *Psychiatry Research: Neuroimaging*, vol. 122, no. 2, pp. 69–87, 2003.

[42] N. C. Andreasen and R. Pierson, “The role of the cerebellum in schizophrenia,” *Biological psychiatry*, vol. 64, no. 2, pp. 81–88, 2008.

[43] H. Picard, I. Amado, S. Mouchet-Mages, J.-P. Olié, and M.-O. Krebs, “The role of the cerebellum in schizophrenia: an update of clinical, cognitive, and functional evidences,” *Schizophrenia bulletin*, vol. 34, no. 1, pp. 155–172, 2008.

[44] J. M. Segall, E. A. Allen, R. E. Jung, E. B. Erhardt, S. K. Arja, K. Kiehl, and V. D. Calhoun, “Correspondence between structure and function in the human brain at rest,” *Frontiers in neuroinformatics*, vol. 6, p. 10, 2012.

[45] Dar, S. U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., & Cukur, T. (2019). Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE transactions on medical imaging, 38(10), 2375-2388.

[46] Zhan, B., Li, D., Wu, X., Zhou, J., & Wang, Y. (2021). Multi-modal MRI image synthesis via GAN with multi-scale gate mergence. IEEE Journal of Biomedical and Health Informatics, 26(1), 17-26.

[47] Kalantar, R., Messiou, C., Winfield, J. M., Renn, A., Latifoltojar, A., Downey, K., ... & Blackledge, M. D. (2021). CT-based pelvic T1-weighted MR image synthesis using UNet, UNet++ and cycle-consistent generative adversarial network (Cycle-GAN). Frontiers in Oncology, 11, 665807.

[48] Cao, B., Zhang, H., Wang, N., Gao, X., & Shen, D. (2020, April). Auto-GAN: self-supervised collaborative learning for medical image synthesis. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 10486-10493).

[49] Chen, J., He, Y., Frey, E. C., Li, Y., & Du, Y. (2021). Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468.

[50] Barhoumi, Yassine, Nidhal C. Bouaynaya, and Ghulam Rasool. "Efficient scopeformer: Towards scalable and rich feature extraction for intracranial hemorrhage detection." IEEE Access (2023).

[51] Meng, Xing, Armin Iraji, Zening Fu, Peter Kochunov, Aysenil Belger, Judy M. Ford, Sara McEwen et al. "Multi-model order spatially constrained ICA reveals highly replicable group differences and consistent predictive results from resting data: A large N fMRI schizophrenia study." NeuroImage: Clinical 38 (2023): 103434.

Gray Matters: ViT-GAN Framework for Identifying Schizophrenia Biomarkers Linking Structural MRI and Functional Connectivity

Yuda Bi*, Anees Abrol*, Sihan Jia*, Jing Sui*, Vince D. Calhoun*
*Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory, Atlanta, GA-30303, USA

Abstract—Brain disorders are often associated with changes in brain structure and function, where functional changes may be due to underlying structural variations. Gray matter (GM) volume segmentation from 3D structural MRI offers vital structural information for brain disorders like schizophrenia, as it encompasses essential brain tissues such as neuronal cell bodies, dendrites, and synapses, which are crucial for neural signal processing and transmission; changes in GM volume can thus indicate alterations in these tissues, reflecting underlying pathological conditions. In addition, the use of the ICA algorithm to transform high-dimensional fMRI data into functional network connectivity (FNC) matrices serves as an effective carrier of functional information. In our study, we introduce a new generative deep learning architecture, the conditional efficient vision transformer generative adversarial network (cEViT-GAN), which adeptly generates FNC matrices conditioned on GM to facilitate the exploration of potential connections between brain structure and function. We developed a new, lightweight self-attention mechanism for our ViT-based generator, enhancing the generation of refined attention maps critical for identifying structural biomarkers based on GM. Our approach not only generates high quality FNC matrices with a Pearson correlation of 0.74 compared to real FNC data, but also uses attention map technology to identify potential biomarkers in GM structure that could lead to functional abnormalities in schizophrenia patients. Visualization experiments within our study have highlighted these structural biomarkers, including the medial prefrontal cortex (mPFC), dorsolateral prefrontal cortex (DL-PFC), and cerebellum. In addition, through cross-domain analysis comparing generated and real FNC matrices, we have identified functional connections with the highest correlations to structural information, further validating the structure-function connections. This comprehensive analysis helps to understand the intricate relationship between brain structure and its functional manifestations, providing a more refined insight into the neurobiological research of schizophrenia.

I. INTRODUCTION

The complex relationship between the brain's structural and functional properties is a critical area of research, especially for understanding brain health and disorders, which significantly impact human quality of life. Schizophrenia is a profoundly serious and complex brain disorder with a wide range of symptoms and manifestations. Research suggests that structural changes in the brain, such as changes in the medial prefrontal cortex and enlargement of the lateral ventricles, may occur in the early stages of schizophrenia [39]. Whether these structural changes are causative factors for the functional abnormalities observed in patients with schizophrenia remains a topic worthy of investigation. Structural MRI (sMRI) and functional MRI (fMRI) are essential

neuroimaging tools that offer unique insights into the brain's physiology in health and disease [1], [2], [3]. sMRI provides high-resolution images crucial for detecting morphological changes in gray matter (GM), such as cortical atrophy and hippocampal shrinkage, which are key indicators of neurological conditions like Alzheimer's disease [4], [5]. In contrast, fMRI captures brain activity patterns, identifying regions active during specific tasks or states, and is vital for studying disorders like schizophrenia or autism where these patterns are often disrupted [6]. The interaction between the brain's structural changes and functional abnormalities is intricate and bidirectional. Structural alterations can lead to functional impairments, affecting cognitive processes and behavior. Despite recognition of this connection, it remains not fully understood due to individual differences and the influences of genetic and environmental factors [7].

The rapid advancement of artificial intelligence, particularly deep learning, offers promising potential in neuroimaging and diagnosis of brain diseases. Data fusion techniques in deep learning have been employed to explore the correlations between structural and functional brain imaging [8]. However, most research in this area has primarily focused on using multimodal information for predicting behavioral outcomes [9] or diagnosing brain disorders [10]. These studies highlight the increased accuracy in detecting brain disorders using multimodal techniques in neuroscience, demonstrating that information from different imaging modalities can be effectively integrated. However, the specific processes by which structural and functional modalities correlate, particularly in the context of specific brain disorders, remain poorly understood. This knowledge gap hinders a better understanding of how structural and functional aspects of the brain influence each other. Developing methods to identify common information shared between structural and functional imaging in brain disorders is critical, yet challenging, as it involves unraveling complex interactions that are not fully explained by current technologies. Nevertheless, existing evidence suggests a significant correlation between inter-subject variations in brain structural networks and those observed in resting-state fMRI networks [11]. This indicates an underexplored yet potentially fruitful area of research in understanding the complex interplay between the brain's structure and function.

Generative deep learning models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), have shown significant success in neuroimaging applications. They excel at simulating neuroimaging data, detecting disease-specific patterns, and enabling modality translation, such as converting T1-weighted to T2-weighted MRI scans [45], [46]. This proficiency suggests the existence of a potential unifying biological or structural principle in brain imaging, bridging the gap between different imaging techniques [12], [13], [14], [15], [16]. The differing techniques of these imaging modalities might intersect at a common point of neural information within the brain, which argues for a unified strategy to gain insights into brain function. However, the integration and synthesis of GM and fMRI data, particularly in deep learning applications, remain comparatively underexplored [17], [18]. This contrasts with the substantial advancements in GAN models that predominantly focus on medical image synthesis between different modalities [47], [48]. fMRI data is typically high-dimensional, posing greater complexity in data fusion and generation. As a result, converting high-dimensional fMRI into an FNC matrix using ICA algorithms for translation between structural and functional images has not been previously explored and is the focus of our investigation.

The emergence of the vision transformer (ViT) marks a significant shift in computer vision, particularly within the realm of neuroimaging [19]. This innovative model diverges from

traditional convolutional neural networks (CNNs) by incorporating mechanisms initially crafted for language processing, such as the self-attention mechanism, enabling a more nuanced and holistic image analysis [20]. ViT's architecture, which dissects images into patches for processing through multiple transformer layers, allows for an in-depth analysis, independent of an image segment's spatial location. However, its computational demands, particularly its $O(N^2)$ complexity, pose challenges, propelling the quest for more streamlined architectures [21].

As researchers seek to enhance the efficiency of ViT through methods like architecture pruning and knowledge distillation without compromising performance [22], [23], the value of ViT in neuroimaging becomes increasingly apparent. The need for precision in brain scans calls for advanced, yet efficient, models, making ViT an attractive option due to its superior interpretability facilitated by attention mechanisms. Notably, ViT has been shown to outperform conventional CNNs in handling complex medical imaging datasets, which raises the possibility that ViT could replace CNNs as the foundational architecture in GANs. Such a shift could utilize ViT's detailed representational capabilities, which could be particularly advantageous for complex neuroscientific studies like attention mapping. By capitalizing on ViT's precision and computational efficiency, it could pave the way for significant breakthroughs in detecting and visualizing specific biomarkers within brain regions, thereby enhancing our understanding of the links between structural and functional neuroimaging. However, for high-dimensional 3D GM images, how to reduce the size of patches to obtain more detailed and accurate attention maps without significantly increasing the computational complexity of the ViT self-attention module is an important issue addressed in this paper.

To solve the problem of structural to functional brain image synthesis, this paper describes 1) the creation of a new conditional GAN model, which is called cEViT-GAN, that can generate functional connectivity matrices from GM data. As the generator and discriminator, an efficient ViT model is used. Since we utilize the attention maps generated by the ViT to identify potential biomarkers of schizophrenia in the brain's structure, we set smaller patches. However, our model remains capable of efficient training and operation. 2) In contrast to conventional self-attention operations, we select a block-wise self-attention layer that significantly reduces the computational cost without compromising performance. The mechanism for block-wise self-attention is versatile. The model accentuates regional relationships by employing self-attention operations in each block, thereby capturing localized patterns and interactions. In contrast, when inter-block self-attention is enabled, it ensures that long-term dependencies across the larger structure are not neglected. This dual strategy is ideal for identifying brain biomarkers from GM data. Focusing on specific regional relationships that indicate certain conditions or abnormalities is essential, but it is also necessary to analyze the entire brain image to completely comprehend and diagnose the issue. 3) To enhance training efficiency, we employ a pretrained ViT patch embedding layer, which was derived from an upstream task of diagnosing schizophrenia using the same model architecture. This allows the GAN model's generator to effectively extract already learned features, thereby improving the training efficiency. 4) In addition, our GAN model has the potential to be used as a biomarker identification tool for identifying the structural and functional connections of the human brain, particularly for various brain diseases such as schizophrenia.

II. RELATED WORKS

Generative adversarial networks (GANs), initially proposed by [24] and extended by [25], have significantly advanced as essential AI tools in various generative tasks. These tasks include

image and signal generation [26], as well as text-to-image and image-to-image synthesis [27]. In medical imaging, GANs play a crucial role in super-resolution, where they enhance image clarity and detail, and in the generation of synthetic images. The creation of these synthetic images is fundamental for data augmentation, training simulations, and the provision of enhanced diagnostic insights without the need for additional radiation exposure or patient involvement [28] [15]. Traditionally, GANs have predominantly utilized CNNs for both the generator and discriminator. However, the emergence of ViTs has led researchers to investigate more efficient architectures for ViT-based GAN models [29] [30]. ViTs are particularly effective in brain imaging, excelling at capturing comprehensive brain patterns, thus ensuring a more complete representation and superior feature extraction [49], [50]. This capability is especially beneficial in recognizing complex neural structures, surpassing the performance of CNNs. For instance, [31] introduced a pre-trained ViT model for classifying brain tumors, addressing the limitations of CNNs that tend to focus predominantly on minute pixel variations. Additionally, [32] demonstrated an enhanced ViT architecture capable of utilizing both structural and functional MRI data for predicting various stages of Alzheimer's disease. Furthermore, the integration of ViT and GAN has emerged as a novel trend in medical imaging. An example of this is the study by Zhao et al. [33], who developed a swin transformer-based GAN model [34] aimed at effective reconstruction of high-resolution MRI images.

In the domain of medical image synthesis, the focus has been on generating images across different modalities, such as CT, MRI, PET, and others. Dalmaz et al. [35] created a new GAN model that combines CNNs with transformer blocks. This model makes it much easier to make medical images that are similar and work better. However, we were hardly able to find related works that corresponded to MRI structural and functional image synthesis, besides our previous works, which synthesized FNC data from given sMRI and achieved a high correlation between real FNC and generated FNC data [36]. However, our previous use of a basic ViT-based GAN architecture was time-consuming and did not include the generation of structural biomarkers, which was a significant shortcoming.

III. METHODS

Our methodology is an innovative combination of deep learning architectures that address the complex problem of generating functional neural connectivity (FNC) maps from 3D GM data. Our model cEViT-GAN was designed to learn and generate high-fidelity FNC representations. We propose an efficient block-wise self-attention technique to avoid the significant computational overhead typically associated with ViT's processing of small image patches. This personalized strategy preserves ViT's tremendous feature extraction capabilities while maintaining computational efficiency, allowing the model to handle the large amounts of data associated with GM. We enhance our methods by superimposing the attentional weights from each layer of the ViT encoders onto the spatial information of 3D GM images, thereby aiding in the creation of sophisticated attention maps that not only reflect activations but also differentiate brain patterns between schizophrenia (SZ) and healthy control (HC) participants. Our technique paves the way for more insightful neuroimaging studies, potentially aiding early diagnosis and intervention efforts for mental health problems, by providing a visual and quantitative differentiation between groups.

A. Generative Adversarial Networks

Integrating generative adversarial networks (GANs) into the domain of medical imaging

necessitates a nuanced understanding of their loss functions. For our specific application of synthesizing FNC maps from 3D GM data, we construct a composite loss function that ensures the generation of realistic and medically informative images. The total loss $\mathcal{L}_{total}$ of our GAN framework is a weighted sum of four components:

$$\mathcal{L}_{total} = \mathcal{L}_G + \mathcal{L}_D + \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{corr}, \quad (1)$$

where $\mathcal{L}_G$ denotes the generator loss, $\mathcal{L}_D$ the discriminator loss, $\mathcal{L}_{MSE}$ the mean squared error loss, and $\mathcal{L}_{corr}$ the correlation loss. The terms λ_1 and λ_2 are hyperparameters that balance the contribution of the MSE loss and the correlation loss, respectively.

The generator loss $\mathcal{L}_G$ is defined as:

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log D(G(\mathbf{z}))], \quad (2)$$

where G is the generator, D is the discriminator, and $\mathbf{z}$ is a point sampled from the generator's input noise distribution $p_{\mathbf{z}}(\mathbf{z})$. The discriminator loss $\mathcal{L}_D$ is formulated as:

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))], \quad (3)$$

where $\mathbf{x}$ represents real data samples from the distribution $p_{data}(\mathbf{x})$. The mean squared error loss $\mathcal{L}_{MSE}$ is incorporated to penalize the pixel-wise differences between the generated and real images, thus preserving the structural integrity of the FNC maps:

$$\mathcal{L}_{MSE} = \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\| G(\mathbf{z}) - \mathbf{x}\|^2\right]. \quad (4)$$

The innovation in our approach is embodied by the correlation loss $\mathcal{L}_{corr}$, which ensures that the statistical dependencies between regions in the generated FNC maps are reflective of the true data. This is crucial for maintaining the biological fidelity of the neural connectivity patterns. The correlation loss is defined as:

$$\mathcal{L}_{corr} = 1 - \frac{\mathbb{E}\left[\left(\mathbf{x} - \mu_{\mathbf{x}}\right)\cdot\left(G(\mathbf{z}) - \mu_{G(\mathbf{z})}\right)\right]}{\sigma_{\mathbf{x}}\sigma_{G(\mathbf{z})}}, \quad (5)$$

where μ and σ denote the mean and standard deviation, respectively. This loss encourages the generated maps to have a correlation structure similar to that of the real FNC maps.

Our GAN architecture also incorporates a conditional input, whereby the generator receives both a sample of noise z and a label indicating the class (SZ or HC). This guides the generator towards producing FNC maps that are not only realistic but also correctly aligned with the specified condition:

$$G(\mathbf{z}|y) \quad \text{where} \quad y \in \{\text{SZ}, \text{HC}\}. \quad (6)$$

In essence, by carefully crafting the loss function and incorporating conditionality, our method aims to drive the

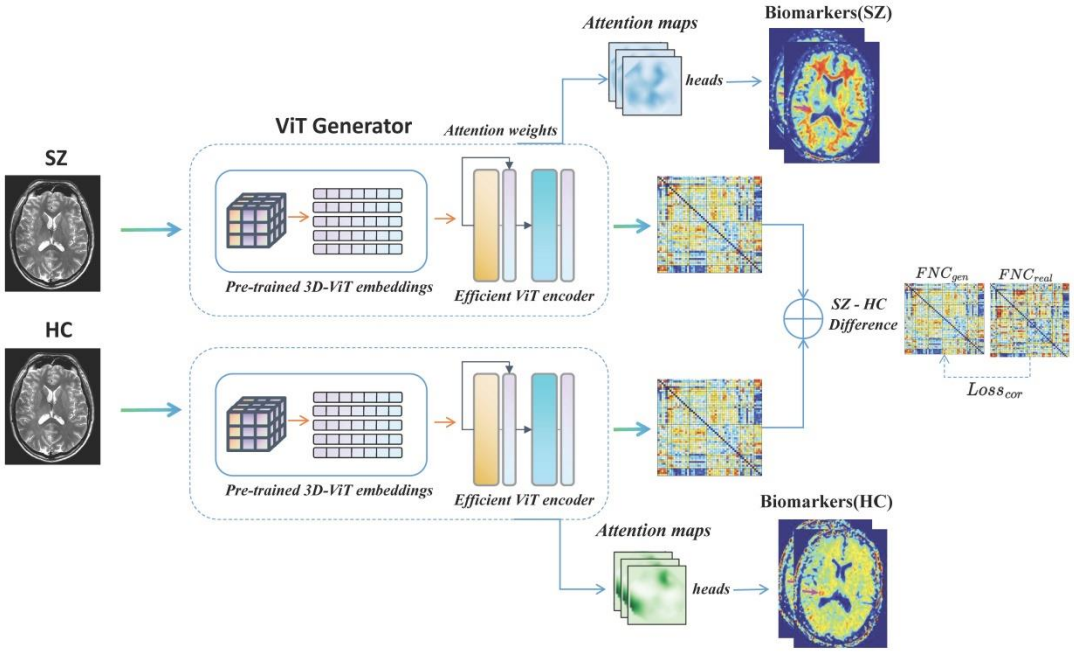

Fig. 1. The proposed methodology involves the analysis of brain MRI scans, specifically those labeled as SZ and HC. The objective is to generate group difference FNC data by utilizing an well-trained efficient generator from the cEViT-GAN framework. Additionally, attention weights are extracted from the ViT encoder to obtain 3D GM attention maps for the different groups. We then apply this approach to identify biomarkers associated with schizophrenia.

GAN towards producing medically valuable outputs.

B. Vision Transformer

The vision transformer (ViT) innovatively adapts transformer mechanisms, originally designed for natural language processing, to computer vision by treating image patches as a sequence of tokens and applying self-attention mechanisms to capture global dependencies within the image.

1) Pre-trained 3D Patch Embedding: The utility of pretrained models in deep learning is unparalleled, particularly in domains where data is scarce or where training from scratch is computationally prohibitive. Leveraging a pre-trained 3D ViT model, our generator benefits from an advanced starting point. This model, initially trained on upstream tasks such as the classification of SZ and HC from GM, has already learned a rich hierarchy of features that are highly relevant to our target domain. The pre-trained model forms the cornerstone of our generator's architecture. Specifically, for the patch embedding process, we utilize the pre-trained embeddings, denoted as:

$$\mathbf{E} = \mathrm{Pre}\left(\left[\mathbf{P}_1; \mathbf{P}_2; \ldots; \mathbf{P}_N\right]\right), \qquad (10)$$

where $\mathbf{P}_i$ corresponds to the flattened vector of the *i*-th 3D patch, and N is the number of non-overlapping 3D patches extracted from the GM input. The function Pre (·) encapsulates the process of obtaining the embedded representations using the pre-trained ViT model.

These pre-trained patch embeddings already encode the spatial hierarchies learned from the upstream classification task, providing a richly structured feature space that is finetuned for the generator:

$$\mathbf{E}^{*} = \mathbf{E} + \mathbf{E}_{pos}, \qquad (11)$$

where $\mathbf{E}^{*}$ represents the embeddings that will be utilized in the transformer encoder, and $\mathbf{E}_{pos}$ is the positional encoding added to the pre-trained embeddings.

These embeddings serve as the input to the ViT encoder, which comprises multiple layers of multi-headed self-attention and feed-forward networks:

$$\mathbf{T} = \mathrm{ViTEncoder}(\mathbf{E}), \qquad (8)$$

where **T** denotes the sequence of transformer encoder outputs corresponding to each patch embedding. Subsequently, each token produced by the ViT encoder is passed through a multilayer perceptron (MLP) network. This MLP is designed to reconstruct the small patches of the generated FNC matrix, transforming the abstract representations learned by the ViT into spatially structured outputs:

$$\mathbf{FNC}_{pa_i} = \mathrm{MLP}\left(\mathbf{T}_i\right). \qquad (9)$$

The collection of FNC patches $\mathbf{FNC}_{pa_i}$ is then reassembled to form the complete FNC map, which serves as the generator's final output:

$$\mathbf{FNC}_{gen} = \mathrm{Reassemble}\left(\left\{\mathbf{FNC}_{pa_1}, \mathbf{FNC}_{pa_2}, \ldots, \mathbf{FNC}_{pa_N}\right\}\right).$$

For the discriminator, the 3D ViT discerns between the real and generated FNC maps, employing a similar patch-based approach to extract features and perform classification. The discriminator's role is to evaluate the authenticity and quality of the generated FNC maps, guiding the generator through the adversarial training process to produce outputs that are increasingly indistinguishable from the real FNC maps derived from GM data. By integrating the ViT model into both the generator and discriminator of our GAN, we harness its potent capacity for capturing intricate patterns and dependencies within the complex data structure of three-dimensional brain imaging.

2) Block-wised Multi-head Self-attention: Incorporating the block-wise multi-head self-attention (BMHSA) [37] mechanism into our model optimizes computing efficiency while keeping the delicate features required for high-resolution biomarker detection from 3D GM data. We used BMHSA in vision tasks because of its excellent performance in dealing with long-text in NLP tasks. BMHSA partitions the collection of 3D GM patch embeddings into smaller, computationally efficient chunks, facilitating focused self-attention within these subdivisions to handle the small patch sizes essential for maintaining resolution in biomarker analysis of 3D GM data. Within each block, BMHSA operates by computing self-attention independently, which drastically reduces the overall computational load compared to traditional methods. Mathematically, the self-attention within a block *b* can be expressed as:

$$\text{Attn}\left(Q_b, K_b, V_b\right) = \text{softmax}\left(\frac{Q_b K_b^T}{\sqrt{d_k}}\right) V_b, \qquad (13)$$

where Q_b, K_b, and V_b are the queries, keys, and values for the block *b*, and d_k represents the scaling factor for the dot products within the softmax function to ensure numerical stability.

Leveraging the concept of multi-head attention, BMHSA allows the model to concurrently attend to different representational subspaces and positions within each block, formulated as:

$$\text{BMHSA}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O, \qquad (14)$$

$$\text{where head}_i = \text{Attn}\left(QW_i^Q, KW_i^K, VW_i^V\right), \qquad (15)$$

with each W_i^Q, W_i^K, and W_i^V denoting the respective parameter matrices for each attention head *i*, and W^O being the output linear transformation matrix.

The BMHSA approach ensures the emphasis of intra-block (regional) relationships while facilitating the preservation of inter-block (long-range) dependencies. These long-range dependencies are crucial for the analysis of structural brain images, as they allow the model to piece together localized information to form a comprehensive understanding of the brain's structure:

$$\mathbf{T} = \text{Concat}\left(\text{BMHSA}\left(\mathbf{E}_1\right), \ldots, \text{BMHSA}\left(\mathbf{E}_B\right)\right) + \mathbf{E}_{pos}, \qquad (16)$$

In this equation, **T** is the output of all the transformer encoder layers put together. It includes both detailed and general information about the brain's structure. The $\mathbf{E}_B$ terms show the embeddings from each block, and the $\mathbf{E}_{pos}$ terms show the positional encodings that are needed to keep the 3D MRI data's natural spatial relationships.

BMHSA Complexity Analysis: By employing the block-wise multi-head self-attention (BMHSA) mechanism, our model achieves significant reductions in computational costs while successfully generating high-resolution attention maps. Traditional self-attention mechanisms, such as those used in ViT models, exhibit a computational complexity that scales quadratically with the length of the sequence *n*. This complexity is expressed as $O\left(n^2 \cdot d\right)$, where *d* is the dimensionality of the attention heads. For long sequences, this scaling becomes computationally prohibitive.

BMHSA addresses this issue by partitioning the input sequence into smaller, fixed-size blocks, each of length *k*. Within each block, self-attention is computed independently, leading to a complexity of $O\left(k^2 \cdot d\right)$ per block. If the input sequence is divided into *m* such blocks, with the total sequence length *n* being equal to $m \times k$, the initial thought would be to express the overall complexity as the sum across all blocks, leading to $O\left(m \cdot k^2 \cdot d\right)$.

However, a more accurate representation of BMHSA's complexity takes into account the parallelizability of these block computations. Since each block's computation is independent, the per-block complexity of $O\left(k^2 \cdot d\right)$ remains, but the computations across different blocks can be performed in parallel. Therefore, the overall computational load does not directly scale with the number of blocks *m*.

Thus, the total computational complexity of BMHSA can be more accurately described as:

$$O\left(k^2 \cdot d\right) \times \text{ parallelization factor,}$$

In conclusion, by judiciously choosing an appropriate block size *k*, BMHSA effectively balances the trade-off between manageable computational costs and the granularity of attention required for detailed analysis in tasks such as high resolution biomarker detection from 3D GM data.

C. cEViT-GAN Architecture

The cEViT-GAN architecture, uniquely designed for analyzing 3D GM data and synthesizing FNC maps, stands out in the field of medical image processing by employing a purely self-attention mechanism instead of standard convolutional techniques. This purely ViT-based approach, in contrast to traditional CNN-based GAN architectures, as detailed in Table 1 which outlines the various layers and functions of our cEViT-GAN model. Figure 2 depicts the pipeline and overall architecture of cEViT-GAN.

Generator Architecture: The generator begins by taking small 3D GM patches, labeled as either SZ or HC . These patches are initially processed through pre-trained 3D embedding layers, utilizing the pre-trained ViT model to capitalize on its extensive feature extraction capabilities from GM data. The data then passes through BMHSA layers, which are crucial for efficient feature extraction and computational load management. The final stage involves MLPs reconstructing the FNC maps from these features, converting transformer outputs into spatially structured FNC patches, which are then assembled into a complete FNC map.

Discriminator Architecture: The discriminator's design features a pure 2D ViT that starts by segmenting FNC maps into patches and processing them through the ViT encoder, effectively discerning patterns to classify the input and produce a probability score indicating the authenticity of the FNC map, a crucial feedback mechanism for the adversarial training of the generator to create accurate and realistic FNC maps.

IV. EXPERIMENT

This section will describe the process of experimental setup, including the datasets and preprocessing, the training and testing of the models, the establishment of baselines, the implementation of the cEViT-GANs, and the experimental design to assess the structural and functional aspects of the brain.

A. Experimental Setups

1) Datasets: In our study, we utilized two comprehensive datasets pertinent to clinical schizophrenia research. Dataset 1 amalgamated data from three distinct studies: fBIRN (Functional Imaging Biomedical Informatics Research Network) across seven sites, MPRC (Maryland Psychiatric Research Center) spanning three sites, and COBRE (Center for Biomedical Research Excellence) at a single site. This aggregation culminated in a total of 827 participants, comprising 477 control subjects (average age: 38.76 ± 13.39, encompassing 213 females and 264

males) and 350 individuals diagnosed with schizophrenia (average age: 38.70 ± 13.14, including 96 females and 254 males). The fBIRN dataset was acquired using uniform resting-state fMRI (rsfMRI) parameters across all sites. We used a standard gradient echo-planar imaging (EPI) sequence with a repetition time (TR) of 2000 ms and an echo time (TE) of 30 ms. The voxels were 3.4375 × 3.4375 × 4 mm in size, and the field of view (FOV) was 220 × 220 mm. The data was captured using six Siemens Tim Trio 3-Tesla scanners and one General Electric Discovery MR750 3.0 Tesla scanner. In the COBRE segment, rsfMRI images were also taken using a standard EPI sequence, but with a slightly different TR/TE of 2000/29 ms and voxel sizes of 3.75 × 3.75 × 4.5 mm, within a field of view (FOV) of 240 × 240 mm, using a 3-Tesla Siemens Tim Trio scanner. The MPRC dataset was gathered using a trio of distinct 3-Tesla Siemens scanners, namely the Siemens Allegra, Trio, and Tim Trio.

Dataset 2 contained a total of 815 subjects, collected from several Chinese hospitals, including 326 subjects (age: 29.81 ± 8.68, females: 167, males: 159) of typical controls and 489 SZ individuals (age: 28.98 ± 7.63, females: 229, males: 260). The subjects were Chinese ethnic Han groups. The dataset was recruited from seven sites in China with the same recruitment criterion, including Peking University Sixth Hospital; Beijing Huilongguan Hospital; Xinxiang Hospital Simens; Xinxiang HospitalGE; Xijing Hospital; Renmin Hospital of Wuhan University; Zhumadian Psychiatric Hospital [51]. The resting-state fMRI data were collected with the following three different types of scanners across the seven sites: 3.0 Tesla Siemens Tim Trio Scanner, 3.0 T Siemens Verio Scanner, and 3.0 T Signa HDx GE Scanner (TR/TE = 2000/30 ms, voxel spacing size = 3 × 3 × 3 mm, FOV = 220 × 220 mm, and 480/360 volumes). Subjects were instructed to relax and lie still in the scanner while remaining calm and awake.

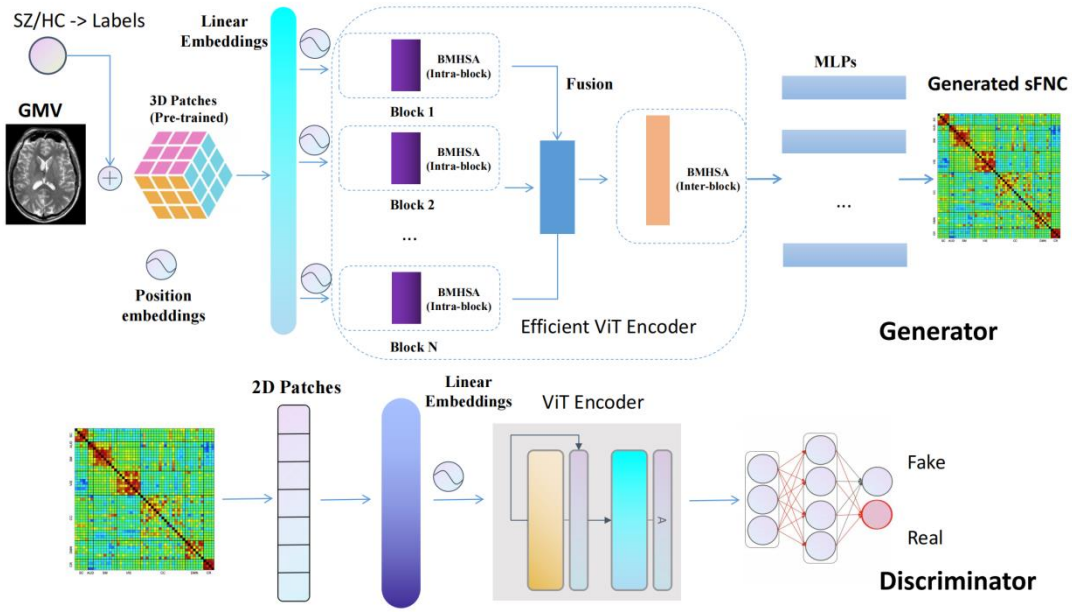

Fig. 2. cEViT-GAN's detailed architecture: The 3D GM and its label (SZ/HC) are input into the cEViT-GAN generator, passing through pre-trained 3D embedding layers and an efficient ViT encoder, followed by MLP outputs to form the FNC. The discriminator architecture is similar to a typical 2D ViT.

2) Pre-processing: To prepare the fMRI data, several critical processes were required: slice

timing correction, realignment, normalization to the EPI template, and smoothing with a 6 mm kernel. Our prior studies contain detailed descriptions of these preprocessing methods. Furthermore, FNC data was obtained using fMRI time series cross-correlation analysis. As spatial priors, a fully automated spatially limited ICA method and the NeuroMark template [38] were utilized. We used a voxel-based morphometry process on the sMRI data to acquire voxel-level GM data.

B. Models

1)Baselines: The primary goal of our comprehensive investigation of the efficacy and performance of several GAN models was to evaluate these models in terms of image-generating capabilities and output quality. The baseline models for comparison were carefully chosen, with special consideration given to their relevance to our pioneering work in synthesizing FNC from GM. While there are no clear previous works for synthesizing FNC, the closest similarity is found in the realm of image synthesis. As a result, we chose GAN models known for their expertise in this field as our baselines.

The first group of baselines includes CNN-based GAN models like Pix2Pix and deep convolutional GAN (DCGAN). The Pix2Pix model, which employs a U-Net generator and a PatchGAN discriminator, is well-known for its ability to solve image-to-image translation problems. The importance of this model in our research arises from its demonstrated ability to generate high-fidelity images from input images, a process that is like our goal of FNC synthesis from GM data. The integration of low and high-level characteristics in the generator by the U-Net architecture improves the detail and quality of the output images. Furthermore, the PatchGAN discriminator focuses on judging the realism of local image patches, which contributes greatly to image sharpness and overall coherence. As a result, these models provide a solid foundation for assessing the potential of GANs in our groundbreaking effort to synthesize FNC from GM. Moreover, we use traditional self-attention-based cViT-GAN as other baselines, which can show the efficiency of our model.

2) cViT-GAN: We employ the traditional cViT-GAN as another baseline for ablation studies, which utilizes conventional self-attention mechanisms. This comparison aims to demonstrate the distinct lightweight advantages of our ViT-encoder that incorporates core blockwise multi-head self-attention (BMHSA) layers. This experiment not only underscores our design's ability to reduce training time and computational complexity but also confirms that our lightweight approach maintains training accuracy despite the simplifications.

3)cEViT-GANs: Our research into novel GAN models resulted in the creation of the cEViT-GAN framework, a revolutionary technique developed exclusively for FNC synthesis. The cEViT-GAN models incorporate cutting-edge approaches, including a pre-training strategy focused on embedding 3D

TABLE I
CEVIT-GAN ARCHITECTURE OVERVIEW

Component	Layer/Function	Description
Generator	Input	Processes 3D GM patches labeled SZ or HC
	Pre-trained 3D Embedding	Utilizes pre-trained ViT model for feature extraction
	BMHSA Layers	Manages computational load, extracts features
	MLPs for Reconstruction	Converts encoder outputs to structured FNC

		patches
	Output Assembly	Assembles patches into complete FNC map
Discriminator	Patch Embedding	Divides FNC maps into patches for processing
	ViT Encoder	Processes embedded patches, extracts features
	Classification Output	MLP that outputs probability of real or generated map

patches and efficient usage of ViT blocks via blockwise self-attention. This novel combination intends to improve picture synthesis quality by combining the strengths of CNNs and transformers.

To optimize cEViT-GAN for our specific needs, we introduced several modifications. The first, cEViT-GAN-b3, consists of three parallel BMHSA blocks without interblock self-attention, prioritizing speed and efficiency while still delivering high quality images. In addition, the cEViT-GAN-b3large includes an interblock self-attention mechanism to enhance the model's ability to capture and integrate complex data patterns for more accurate FNC synthesis. In addition, the cEViT-GAN-b6 uses six parallel BMHSA blocks without inter-block connections to explore the effects of increased parallelism on computational efficiency and image quality. Our most advanced configuration, the cEViT-GAN-b6large, forms the cornerstone of our study and serves as the basis for all visualization and analysis. This model combines multiple BMHSA layers with interblock self-attention, designed to balance precise feature acquisition with efficient processing of 3D GM data. The inclusion of interblock self-attention is critical, as it allows for more effective integration of information across layers, which can lead to more refined and accurate synthesis of FNC from GM data.

E. Experiments Details

1) Pre-training: When testing our ViT-based GAN models, including the baseline and our cViT-GAN variants, we utilize pre-trained 3D linear embeddings. These embeddings are obtained from a previously developed multimodal deep learning model designed for schizophrenia diagnosis and classification [17]. The significance of this pre-training is particularly pronounced for the generator components of our GANs, facilitating an efficient transfer of learned features across medical imaging tasks. This enhances the generalizability and robustness of our models.

The generators in our cEViT-GANs benefit significantly from starting with weights derived from these pre-trained embeddings. This not only accelerates the training process by providing an informed initialization but also improves the models' overall efficiency and effectiveness. The embeddings encapsulate a wealth of features relevant to schizophrenia, enriching the generators with nuanced neuroimaging patterns associated with the condition. Consequently, our GAN models can synthesize FNC images from GM data that are more detailed, and clinically relevant to our neurological focus.

2) Train and Validation: It's worth noting that while all CNN-based GAN models employ a uniform set of parameters and training techniques, ViT-based GANs, including baseline and cEViT-GAN variations, utilize a distinct set. For CNN-based GANs, we use Kaiming initialization for selecting initial weights and the AdamW optimizer for both the generator and the discriminator, setting the learning rate at 1e-3 with a MultistepLR schedule that adjusts at the 20th, 50th, and 150th epochs. Conversely, ViT-based GAN models, which incorporate pre-trained weights for 3D patch embedding in the generator, require a lower learning rate of 1e-4, also with AdamW as the

optimizer and a MultistepLR schedule making adjustments at the 20th, 50th, and 90th epochs. All models are trained with a batch size of 32, and the pre-training stage typically leads to convergence around the 90th epoch.

Using cross-validation on both types of models improves model resilience and dependability. This technique is useful for determining how the models would perform on different data sets, decreasing the danger of over-fitting and assuring generalization. Our training and validation operations are powered by 8 NVIDIA Tesla V100 GPUs, and we utilize PyTorch as our model framework. We adopt parallel and distributed training approaches, specifically using PyTorch's distribution methods, to distribute the training load across multiple GPUs. This strategy not only enhances processing efficiency but also significantly reduces training times. It is particularly beneficial for handling the large volumes of data and complex neural network architectures required in our research. In addition, we use PyTorch's built-in distribution strategies, which use the all-reduce algorithm rather than a parameter server approach. This method efficiently aggregates gradients across multiple GPUs to ensure synchronized updates and optimal training performance.

F. Evaluation Metrics

In our research, we deploy a trio of critical metrics to assess the efficacy of our model in synthesizing FNC patterns. These metrics include the Mean Squared Error (MSE), the Pearson Correlation Coefficient, and the Cosine Similarity. Each of these metrics plays a crucial role in evaluating the precision and reliability of the FNC patterns generated by our model, offering distinct insights into the model's performance and facilitating a comprehensive assessment when compared to authentic FNC data.

1) Mean Squared Error (MSE): MSE is a metric commonly utilized in regression analysis and signal processing. It quantifies the average of the squares of errors, which are the differences between the estimated values and the actual values. In the context of our FNC data, the MSE is calculated as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Here, n denotes the total number of FNC entries, Y_i represents the actual FNC value, and $\hat{Y}_i$ signifies the estimated FNC value produced by the model. A lower MSE value is indicative of superior model performance, signifying a reduced deviation from the true FNC values.

2) Pearson Correlation Coefficient (PCC): The PCC is a measure that quantifies the linear correlation between two datasets. It yields a value within the range of -1 to 1, where 1 denotes a total positive linear correlation, 0 signifies no linear correlation, and -1 indicates a total negative linear correlation. In evaluating our model, this coefficient is defined as:

$$r = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(\hat{Y}_i - \bar{\hat{Y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}}$$

Where $\bar{Y}$ and $\bar{\hat{Y}}$ are the mean values of the actual and estimated FNCs, respectively. A

higher absolute value of this coefficient implies a stronger correlation between the generated FNCs and the real data.

3) Cosine Similarity: Cosine Similarity is a metric employed to ascertain the similarity between two vectors, irrespective of their magnitude, and is especially pertinent in high-dimensional spaces. The cosine similarity between the actual and the model-generated FNC vectors is computed as:

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n} Y_i \times \hat{Y}_i}{\sqrt{\sum_{i=1}^{n} Y_i^2} \times \sqrt{\sum_{i=1}^{n} \hat{Y}_i^2}}$$

In this formula, the numerator represents the dot product of the actual and estimated FNC vectors, while the denominator is the product of the Euclidean norms of these vectors.

Together, these metrics provide a strong and flexible way to check how accurate and similar the FNC patterns our model creates are to real FNC data. This gives us important information about how well the model can copy complex neural connectivity patterns.

G. Visualizations

We performed intense brain anatomical and functional visualization based on self-attention operations in two phases. The first step is to use attention weights on original brain maps to find any biomarkers in GM data while also making a matching FNC. The second step is to compare the made FNC with the real FNC, which could help find functional biomarkers for SZ disease.

1) MRI Attention Maps: To extract attention weights, we used a rollout method, concatenating weights from each block from blockwise multihead self-attention and superimposing the weights of interblock self-attention onto the averaged block weights, producing a comprehensive attention map useful for detecting potential biomarkers in GM data. High attention weights help us identify regions of interest that may be connected with various neurological diseases, such as SZ. The attention map serves as a guide, showing the most important sections of the GM data. Researchers can get insights into the fundamental mechanisms of SZ and potentially other neurological illnesses by better understanding these areas. The incorporation of attention weights into GM data analysis represents a significant improvement in neuroimaging and the research of brain diseases. Fortunately, analyzing the group difference (HZ-HC) attention map allowed us to identify brain areas strongly associated with SZ that aligned with our existing knowledge. Following on from the analysis of group differences, our model also supports the generation of FNC maps based on individual cases of GM. This capability represents a promising avenue for the exploration of personalized biomarkers, as it allows the adaptation of our approach to individual variations in GM data. By tailoring the analysis to each individual subject, researchers can potentially uncover unique patterns and connections in brain structures that are specific to individual neurological profiles, increasing the precision and relevance of biomarker discovery in conditions such as SZ and other neurological diseases.

2) FNC Maps: Our cEViT-GANs are capable of generating reasonably accurate FNC maps. We generate FNC maps for each subject by conditioning on the GM data from each group (SZ and HC), and then average these maps to assess and validate the accuracy and effectiveness of our model. Further, by analyzing the averaged group differences in FNC maps, we demonstrate that our model can effectively learn these distinctions, using GM as structural input to derive functional differences, thereby substantiating the biological significance of the connection

between structure and function.

V. RESULTS

This section presents the outcomes of the experiments as well as a visualization of the structure and function of the brain using attention maps and FNC biomarkers. Initially, we conducted exhaustive experiments on various baselines and our cEViT-GAN variants to ensure that our model exhibited superior accuracy and robustness.

A. Model Performance

The experimental results shown in Figure 3 highlight the performance differences between our baseline models (Pix2Pix, DCGAN, and cViT-GAN) and our new cEViT-GAN variants. From these results, it is clear that the standard DCGAN, which uses a pure CNN backbone, underperforms in FNC generation, suggesting that pure CNN architectures are not particularly effective for this task. Pix2Pix, a well-known GAN model that adapts both the generator and the discriminator, meets the requirement for high quality image generation to some extent.

The use of a pure ViT backbone, as in cViT-GAN, is advantageous for the extraction of long-range features due to its self-attention mechanism, but results in higher computational costs. Further reduction of the patch size in this context could lead to memory overload. Our proposed cEViT-GAN, especially the cEViT-GAN-b6large model with interblock self-attention, shows excellent performance; however, the inclusion of interblock self-attention increases the training time. The base model of EViT-GAN, such as cEViT-GAN-b6, significantly reduces training complexity without compromising accuracy - maintaining the same level of accuracy as cViT-GAN but with reduced training time. Therefore, as we explore ways to further reduce patch sizes in the future, the use of EViT-GAN could not only reduce training times, but also maintain the quality and accuracy of the generated FNCs while improving the refinement of attention maps.

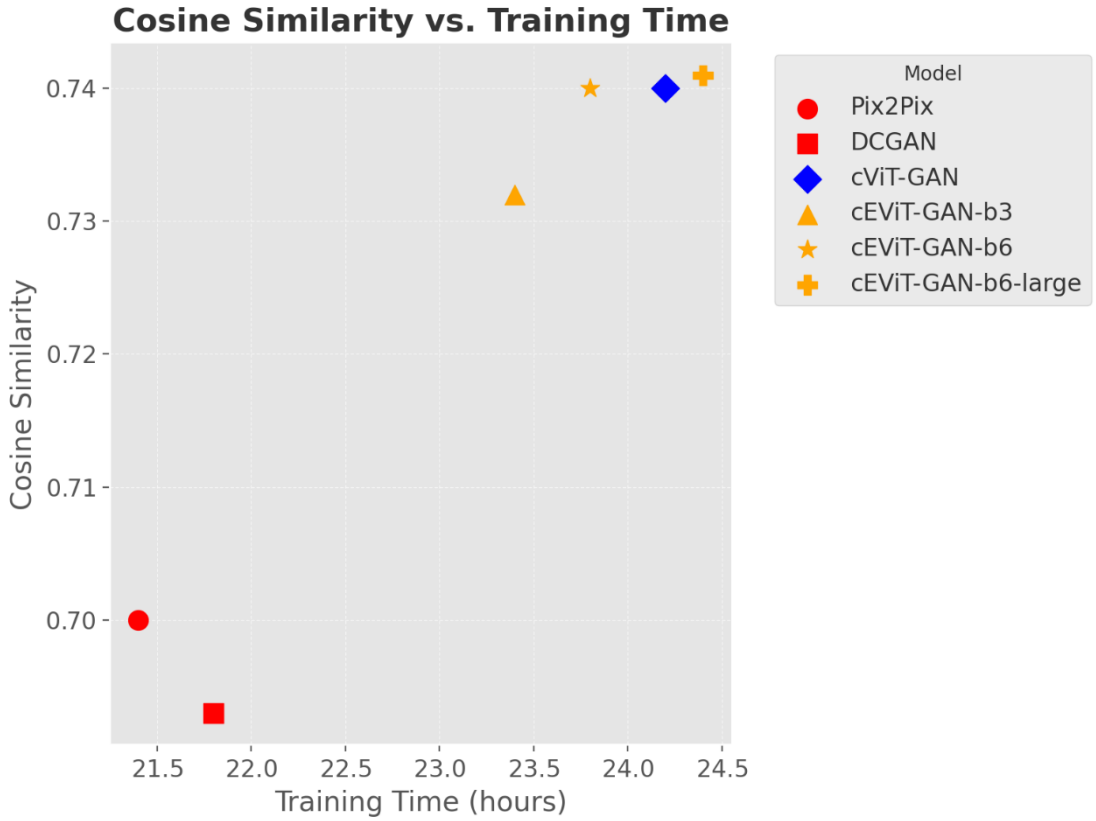

Fig. 3. Comparison of different models in terms of Cosine Similarity and Training Time.

TABLE II
MODEL PERFORMANCE COMPARISON

Model	Backbone	Cosine Similarity	Pearson Correlation	Training Time (hours)
Pix2Pix	CNN	0.7	0.71	21.4
DCGAN	CNN	0.693	0.693	21.8
cViT-GAN	ViT	0.74	0.74	24.2
cEViT-GAN-b3	ViT with BMHSA	0.732	0.731	23.4
cEViT-GAN-b6	ViT with BMHSA	0.74	0.741	23.8
cEViT-GAN-b6-large	ViT with BMHSA	0.741	0.741	24.4

B. MRI Attention Maps

We analyzed attention weights in our cEViT-GAN generator to generate our 3D GM attention maps. We created subject-specific attention maps for each member of our testing set, then tested for group differences using a two-sample t-test. Each voxel in our attention maps represents a t-value from this statistical test. To account for multiple comparisons, we used the false discovery rate (FDR) method with a $q < 0.05$ threshold. This approach accounts for the possibility of type I errors while running several statistical tests. The attention maps that arise emphasize areas with statistically significant changes in activation patterns between groups. Figure 4 shows the attention maps in a three-plane view.

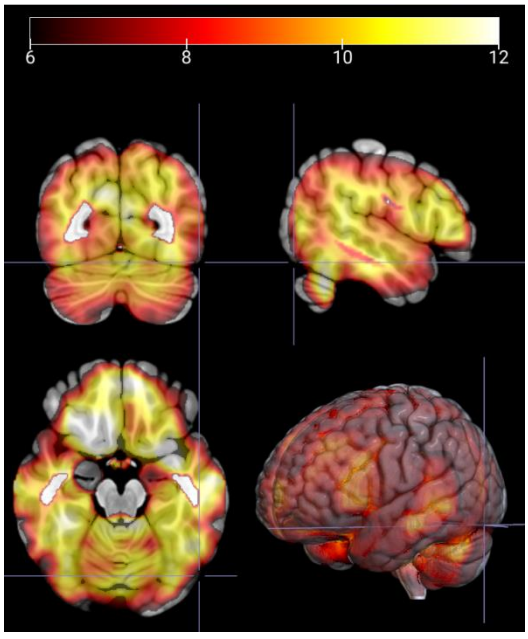

Fig. 4. The 3D MRI attention maps for group difference analysis (SZ vs HC), which indicate the significant ROIs that are strongly associated with schizophrenia disease.

Figure 4 indicates that while generating the related functional outputs, our model prioritized four brain regions: the medial pre-frontal cortex (mPFC), the dorsolateral prefrontal cortex (DL-PFC), the temporal lobe, and the cerebellum. Schizophrenia is a diverse, complex psychiatric

condition that frequently involves dysfunctions in numerous brain circuits. Based on traditional neuroscience and previous knowledge, mFPC is intimately related to executive processes and decision-making, both of which can be affected in schizophrenia [39]. The mPFC is also involved in emotional processing, and abnormalities here can be linked to negative schizophrenia symptoms including apathy and social disengagement [40]. DL-PFC is required for cognitive control and working memory, both of which are frequently impaired in people with schizophrenia. Deficits in this area can contribute to the disorder's hallmarks of disorganized thinking and trouble focusing attention. The superior temporal gyrus, in particular, is connected with auditory processing and language. Temporal lobe dysfunction has been linked to auditory hallucinations and language difficulties seen in schizophrenia patients [41]. Finally, the cerebellum's significance in cognitive processing is now recognized. Cerebellar abnormalities may contribute to cognitive impairments and affective dysregulation in schizophrenia, according to recent research [42], [43].

C. FNC Biomarkers

1) FNC Analysis: In this study, a sophisticated GAN model was employed to generate FNC outputs from a test dataset. Our analysis revealed that the model's output for the whole average FNC exhibited a strong correlation (0.97) with the actual FNC data across all subjects. This is effectively visualized in Figure 5, which compares the model-generated whole average FNC with the genuine FNC data. The ability of our GAN model to replicate FNC from 3D MRI scans of GM with high accuracy can be credited to the identification of neural structures via independent component analysis (ICA). ICA has been known to uncover network-like structures within the resting gray matter, providing insights into the model's capability to replicate these intricate neural patterns. The correlation observed in our model's output with the real data not only validates our approach but also aligns it with previous scientific research in neuroimaging [44], [16], [11].

2) Group Difference Analysis: We also show the produced and real group-difference FNC (HC-SZ). Figure 5 shows a comparison of calculated and actual FNC group differences. Our model can infer group-difference FNC from brain structure with a remarkably high correlation (0.74) especially given brain function contains unique information above and beyond brain structure. Our cEViT-GAN model can identify a strong similarity between the generated group-difference FNC and the real one, and the patterns are those that are know to be implicated in schizophrenia, including subcortical areas. These include connections between the cerebellum and the subcortical (CB-SC), auditory (CB-AUD), somatomotor (CB-SM), visual (CB-VS), cingulo-opercular (CB-CC), default mode (CB-DM), and the cerebellum itself (CB-CB). The synthetic FNC data obtained by GM has a remarkable correlation with real FNC data, with similarities reaching 0.85 in certain subcortical linkages.

This important finding shows that subcortical structures are important for identifying differences between HC and SZ participants and that the cEViT-GAN model has a good performance of showing these important structural-functional relationships. The remarkable similarity in subcortical areas shows that our model is quite good at reproducing complicated, potentially clinically relevant brain patterns. Such capabilities signal new opportunities to improve our understanding of diseases such as schizophrenia, to offer more precise diagnostic measures, and to personalize therapy methods. Furthermore, our model demonstrates that there is a high level of agreement in the difference in values for other pairs of connections, such as cingulo-opercular (CC-CC), somatomotor-default mode (SM-DM), and visual-default mode (VS-DM). We also find moderate parallelism in visual-auditory (VS-AUD) and

cingulo-opercular-somatomotor (CC-SM) pairs. These findings provide greater insight into how the disparities in FNC observed between the HC and SZ groups may be caused by underlying structural issues. The combined insights are critical for developing more refined diagnostic tools and therapy approaches for navigating the complexities of schizophrenia.

3) Cross-domain Analysis: Our FNC matrix cross-domain analysis provides a more detailed view of the link between structural and functional data. The produced and real FNC matrices have a total similarity measure of 0.74, which shows that there is a significant relationship, but the structural data does not fully reflect all functional features. The complicated nature of brain functionality, which cannot be entirely extrapolated from structural imaging, may account for this disparity.

Upon examining the cross-domain correlations, it becomes apparent that the within-domain correlations, such as AUD-AUD, exhibit a remarkably high similarity (0.955). This suggests that the structural data accurately reflects the functional connection of the auditory network. This is supported by strong correlations in subdomains like SC-AUD (0.847) and SC-SM (0.824), which show stable structural-functional alignment in the motor function and sensory processing domains.

The cEViT-GAN model captures the cerebellum's constant functional patterning, as evidenced by its strong intra-domain correlation (CB-CB at 0.821). Cross-domain interactions, like those between the default mode network and the cerebellum (DMN-CB) and the somatomotor and cerebellar regions (SM-CB), have moderate to high correlations. This means that the model can show how different parts of the brain work together. Notably, the lower correlations in coupling between other regions including SC-CB (0.160) and AUD-CB (0.053) show the challenge of mapping functional networks from structural data, especially when there are complicated connections between regions. These areas may indicate distinct functional characteristics or dynamic interconnections that are not readily apparent in GM data. These correlations are specific across domain sizes, from the small 2x2 matrices to the large 17x17 matrices. This makes them useful for checking the authenticity of FNC representations that have been made. Furthermore, it identifies areas where the generative model's performance could be enhanced to more accurately replicate the intricate tapestry of human brain connectivity.

Finally, our findings highlight the benefits and drawbacks of employing cEViT-GAN to replicate FNC matrices using structural data. The model's high fidelity in some domains encourages its use in clinical settings, whereas inequalities in others call for further research into the multidimensional nature of brain structure-function interactions.

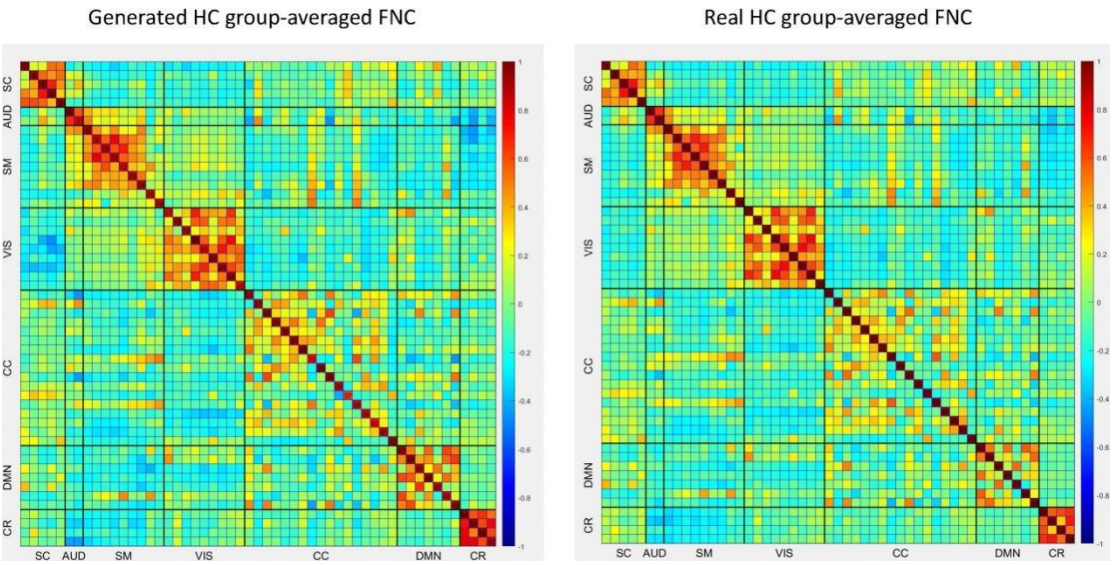

Fig. 5. The generated whole average FNC vs. real whole average FNC

D. Structural-to-functional Connectivity

The identification of biomarkers in SZ by combining structural and functional neuroimaging data tells a captivating story about the disorder's neuropathology. The findings of our investigation show a significant agreement between structural and functional indicators, highlighting the complicated connection between brain structure and function in SZ.

Significant structural sections include the medial frontal cortex (mPFC), dorsolateral prefrontal cortex (DL-PFC), and cerebellum. Similar functional regions show significant changes in connection patterns, particularly in the default mode network (DMN) and auditory and somatomotor activities. This correspondence between structural changes and functional connectivity disturbances allows for a more comprehensive understanding of SZ pathophysiology. For example, functional connectivity disruptions in the mPFC and DL-PFC, which are important for executive functioning and cognitive control, coincide with structural alterations in these areas, contributing to the cognitive and affective dysregulation seen in SZ patients. Both structural and functional findings highlight the importance of the cerebellum in SZ, an area that has been understudied until now. Changes in cerebellar areas correspond structurally with changes in functional connectivity within the cerebellum and its linkages to other brain networks. This shows that the cerebellum may play a role in the larger network dysfunctions that characterize SZ, going beyond its traditional concept of motor control.

Furthermore, the temporal lobe, a region involved in auditory processing, exhibits both structural and functional abnormalities, which correspond to clinical symptoms such as auditory hallucinations, which are common in SZ. This is supported by the significant correlation in functional networks, including the auditory cortex (AUD), which mirrors the anatomical findings. These similarities hint at a more integrated model of SZ in which structural anomalies are not isolated but have a considerable impact on the functional network dynamics. This model supports the idea that SZ is a disorder of "disconnected connection," with the symptoms being caused by the interaction of damage to the structure and problems with the way the network works.

In conclusion, the convergence of structural and functional biomarkers in our work have provided some new insights into our understanding of SZ. It demonstrates the interrelated nature of structural and functional network changes, providing a more comprehensive view of the

disorder's neurobiological roots. We hope our understanding can be further increase by an integrative approach like this, pontentially leading to the development of more effective diagnostic tools and targeted treatment options that are tailored to the personalized nature of SZ.

VI. DISCUSSION AND CONCLUSION

In this study, we introduced the cEViT-GAN, a novel approach that combines GAN with ViT and a new lightweight blockwise multihead self-attention technique. This model effectively generates FNC matrices from brain structural GM data, supporting the neuroscientific and biological perspective that there is a link between brain structure and function.

In particular, in neurological disorders such as schizophrenia, changes in brain function are often due to underlying changes in brain GM structure. By analyzing the results generated on a per-subject basis, our use of attention map technology has enabled the pinpointing of brain structures, such as the medial prefrontal cortex (mPFC), dorsolateral prefrontal cortex (DL-PFC), and cerebellum, that drive functional changes. Furthermore, our model effectively simulates characteristics and changes similar to those found in real FNCs, providing potential evidence that these functional changes originate from specific brain structures. This is particularly evident when comparing generated FNCs with real FNCs.

However, our model has limitations. For example, the conditional generative model, which typically operates under the supervision of a target generative object, is influenced by that target and attempts to replicate its statistical properties. Consequently, the FNC matrices generated by our model are supervised by actual FNC data and are not generated solely on the basis of structural GM data. To isolate the unique information derived from FNC structural data, it is necessary to eliminate the influence of unconditional generation from actual FNC data, a factor not addressed in our experiments.

Despite these limitations, our model represents a pioneering exploration of the use of data-driven 3D structural data to generate high quality FNCs. Our findings on a schizophrenia dataset provide guidance for future work. In the future, we aim to further extend and validate our model to develop a more generalized pipeline that is potentially applicable to a broader range of brain disorders and datasets.

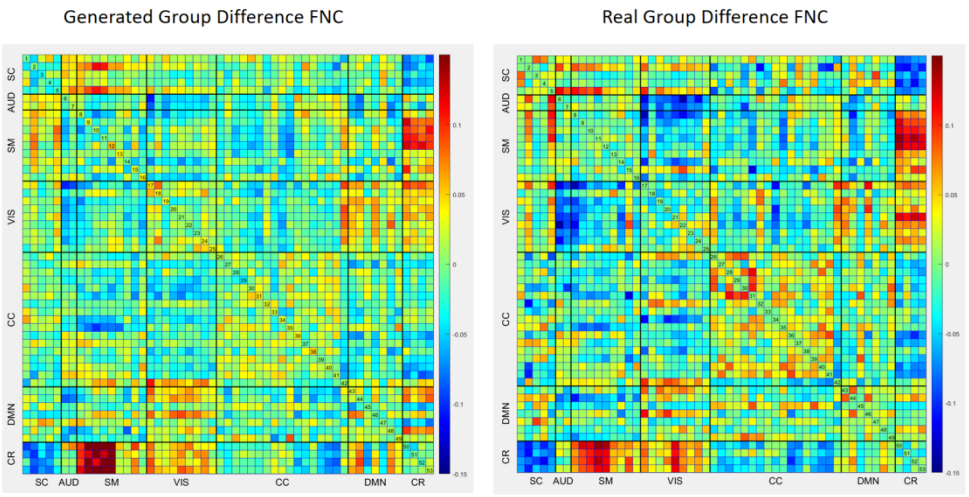

Fig. 6. The generated 'fake' group difference FNC matrix vs. real group difference FNC matrix.

Ethics for data: We are doing secondary analysis of data from public and private repositories. All data were collected under appropriate ethics approval and all subject signed informed consent.

CRediT authorship contribution statement: Yuda Bi: Investigation, Validation, Visualization, and Writing – original draft, Writing – review & editing. Anees Abrol: Formal analysis, Writing – review & editing. Sihan Jia: Writing – review & editing. Jing Sui: Writing – review & editing, Data curation. Vince D. Calhoun: Funding acquisition, Resources, Writing – review & editing.

REFERENCES

[1] G. D. Pearlson and V. Calhoun, "Structural and functional magnetic resonance imaging in psychiatric disorders," *The Canadian Journal of Psychiatry, vol*. 52, no. 3, pp. 158–166, 2007.

[2] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative *et al*., "Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database," *neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[3] U. Khatri and G.-R. Kwon, "Alzheimer's disease diagnosis and biomarker analysis using resting-state functional mri functional brain network with multi-measures features and hippocampal subfield and amygdala volume of structural mri," *Frontiers in aging neuroscience*, vol. 14, p. 818871, 2022.

[4] X. Zhao, C. K. E. Ang, U. R. Acharya, and K. H. Cheong, "Application of artificial intelligence techniques for the detection of alzheimer's disease using structural mri images," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 456–473, 2021.

[5] N. Franzmeier, N. Koutsouleris, T. Benzinger, A. Goate, C. M. Karch, A. M. Fagan, E. McDade, M. Duering, M. Dichgans, J. Levin *et al*., "Predicting sporadic alzheimer's disease progression via inherited alzheimer's disease-informed machine-learning," *Alzheimer's & Dementia*, vol. 16, no. 3, pp. 501–511, 2020.

[6] J. Oh, B.-L. Oh, K.-U. Lee, J.-H. Chae, and K. Yun, "Identifying schizophrenia using structural mri with a deep learning algorithm," *Frontiers in psychiatry*, vol. 11, p. 16, 2020.

[7] C. J. Honey, J.-P. Thivierge, and O. Sporns, "Can structure predict function in the human brain?" *Neuroimage*, vol. 52, no. 3, pp. 766– 776, 2010.

[8] V. D. Calhoun and J. Sui, "Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness," *Biological psychiatry: cognitive neuroscience and neuroimaging*, vol. 1, no. 3, pp. 230–244, 2016.

[9] J. Sui, R. Jiang, J. Bustillo, and V. Calhoun, "Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises," *Biological psychiatry*, vol. 88, no. 11, pp. 818–828, 2020.

[10] B. Rashid and V. Calhoun, "Towards a brain-based predictome of mental illness," *Human brain mapping*, vol. 41, no. 12, pp. 3468–3535, 2020.

[11] N. Luo, J. Sui, A. Abrol, J. Chen, J. A. Turner, E. Damaraju, Z. Fu, L. Fan, D. Lin, C. Zhuo *et al*., "Structural brain architectures match intrinsic functional networks and vary across domains: a study from 15 000+ individuals," *Cerebral Cortex*, vol. 30, no. 10, pp. 5460–5470, 2020.

[12] J. Pan, B. Lei, Y. Shen, Y. Liu, Z. Feng, and S. Wang, "Characterization multimodal connectivity of brain network by hypergraph gan for alzheimer's disease analysis," in *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4*. Springer, 2021, pp. 467–478.

[13] X. Dai, Y. Lei, Y. Fu, W. J. Curran, T. Liu, H. Mao, and X. Yang, "Multimodal mri synthesis using unified generative adversarial networks," *Medical physics*, vol. 47, no. 12, pp. 6343–6354, 2020.

[14] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "Gans for medical image synthesis: An empirical study," *Journal of Imaging*, vol. 9, no. 3, p. 69, 2023.

[15] Y. Liu, A. Chen, H. Shi, S. Huang, W. Zheng, Z. Liu, Q. Zhang, and X. Yang, "Ct synthesis from mri using multi-cycle gan for head-andneck radiation therapy," *Computerized medical imaging and graphics*, vol. 91, p. 101953, 2021.

[16] N. Luo, J. Sui, A. Abrol, D. Lin, J. Chen, V. M. Vergara, Z. Fu, Y. Du, E. Damaraju, Y. Xu *et al*., "Age-related structural and functional variations in 5,967 individuals across the adult lifespan," *Human brain mapping*, vol. 41, no. 7, pp. 1725–1737, 2020.

[17] Y. Bi, A. Abrol, Z. Fu, and V. Calhoun, "Multivit: Multimodal vision transformer for schizophrenia prediction using structural mri and functional network connectivity data," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.

[18] M. A. Azam, K. B. Khan, S. Salahuddin, E. Rehman, S. A. Khan, M. A. Khan, S. Kadry, and A. H. Gandomi, "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Computers in biology and medicine*, vol. 144, p. 105253, 2022.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al*., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking," *arXiv preprint arXiv:2309.02031*, 2023.

[22] Y. Tang, K. Han, Y. Wang, C. Xu, J. Guo, C. Xu, and D. Tao, "Patch slimming for efficient vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 165–12 174.

[23] X. Chen, Q. Cao, Y. Zhong, J. Zhang, S. Gao, and D. Tao, "Dearkd: data-efficient early knowledge distillation for vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 052–12 062.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[25] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401– 4410.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125– 1134.

[28] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. Á. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, “Madgan: Unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction,” *BMC bioinformatics*, vol. 22, no. 2, pp. 1–20, 2021.

[29] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, “Vitgan: Training gans with vision transformers,” *arXiv preprint arXiv:2107.04589*, 2021.

[30] S. Hirose, N. Wada, J. Katto, and H. Sun, “Vit-gan: Using vision transformer as discriminator with adaptive data augmentation,” in *2021 3rd International Conference on Computer Communication and the Internet (ICCCI)*. IEEE, 2021, pp. 185–189.

[31] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, “Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling,” *Current Oncology*, vol. 29, no. 10, pp. 7498– 7511, 2022.

[32] S. Sarraf, A. Sarraf, D. D. DeSouza, J. A. Anderson, M. Kabia, and A. D. N. Initiative, “Ovitad: Optimized vision transformer to predict various stages of alzheimer’s disease using resting-state fmri and structural mri data,” *Brain Sciences*, vol. 13, no. 2, p. 260, 2023.

[33] X. Zhao, T. Yang, B. Li, and X. Zhang, “Swingan: A dual-domain swin transformer-based generative adversarial network for mri reconstruction,” *Computers in Biology and Medicine*, vol. 153, p. 106513, 2023.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *in Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[35] O. Dalmaz, M. Yurt, and T. Çukur, “Resvit: Residual vision transformers for multimodal medical image synthesis,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[36] Y. Bi, A. Abrol, J. Sui, and V. Calhoun, “Cross-modal synthesis of structural mri and functional connectivity networks via conditional vitgans,” *arXiv preprint arXiv:2309.08160*, 2023.

[37] J. Qiu, H. Ma, O. Levy, S. W.-t. Yih, S. Wang, and J. Tang, “Blockwise self-attention for long document understanding,” *arXiv preprint arXiv:1911.02972*, 2019.

[38] Y. Du, Z. Fu, J. Sui, S. Gao, Y. Xing, D. Lin, M. Salman, A. Abrol, M. A. Rahaman, J. Chen *et al*., “Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders,” *NeuroImage: Clinical*, vol. 28, p. 102375, 2020.

[39] X. J. Chai, S. Whitfield-Gabrieli, A. K. Shinn, J. D. Gabrieli, A. Nieto Castañón, J. M. McCarthy, B. M. Cohen, and D. Öngür, “Abnormal medial prefrontal cortex resting-state connectivity in bipolar disorder and schizophrenia,” *Neuropsychopharmacology*, vol. 36, no. 10, pp. 2009– 2017, 2011.

[40] J. H. Callicott, A. Bertolino, V. S. Mattay, F. J. Langheim, J. Duyn, R. Coppola, T. E. Goldberg, and D. R. Weinberger, “Physiological dysfunction of the dorsolateral prefrontal cortex in schizophrenia revisited,” *Cerebral cortex*, vol. 10, no. 11, pp. 1078–1092, 2000.

[41] L. L. Davidson and R. W. Heinrichs, “Quantification of frontal and temporal lobe brain-imaging findings in schizophrenia: a meta-analysis,” *Psychiatry Research: Neuroimaging*, vol. 122, no. 2, pp. 69–87, 2003.

[42] N. C. Andreasen and R. Pierson, “The role of the cerebellum in schizophrenia,” *Biological psychiatry*, vol. 64, no. 2, pp. 81–88, 2008.

[43] H. Picard, I. Amado, S. Mouchet-Mages, J.-P. Olié, and M.-O. Krebs, “The role of the cerebellum in schizophrenia: an update of clinical, cognitive, and functional evidences,” *Schizophrenia bulletin*, vol. 34, no. 1, pp. 155–172, 2008.

[44] J. M. Segall, E. A. Allen, R. E. Jung, E. B. Erhardt, S. K. Arja, K. Kiehl, and V. D. Calhoun, “Correspondence between structure and function in the human brain at rest,” *Frontiers in neuroinformatics*, vol. 6, p. 10, 2012.

[45] Dar, S. U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., & Cukur, T. (2019). Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE transactions on medical imaging, 38(10), 2375-2388.

[46] Zhan, B., Li, D., Wu, X., Zhou, J., & Wang, Y. (2021). Multi-modal MRI image synthesis via GAN with multi-scale gate mergence. IEEE Journal of Biomedical and Health Informatics, 26(1), 17-26.

[47] Kalantar, R., Messiou, C., Winfield, J. M., Renn, A., Latifoltojar, A., Downey, K., ... & Blackledge, M. D. (2021). CT-based pelvic T1-weighted MR image synthesis using UNet, UNet++ and cycle-consistent generative adversarial network (Cycle-GAN). Frontiers in Oncology, 11, 665807.

[48] Cao, B., Zhang, H., Wang, N., Gao, X., & Shen, D. (2020, April). Auto-GAN: self-supervised collaborative learning for medical image synthesis. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 10486-10493).

[49] Chen, J., He, Y., Frey, E. C., Li, Y., & Du, Y. (2021). Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468.

[50] Barhoumi, Yassine, Nidhal C. Bouaynaya, and Ghulam Rasool. "Efficient scopeformer: Towards scalable and rich feature extraction for intracranial hemorrhage detection." IEEE Access (2023).

[51] Meng, Xing, Armin Iraji, Zening Fu, Peter Kochunov, Aysenil Belger, Judy M. Ford, Sara McEwen et al. "Multi-model order spatially constrained ICA reveals highly replicable group differences and consistent predictive results from resting data: A large N fMRI schizophrenia study." NeuroImage: Clinical 38 (2023): 103434.

Data and Code Availability Statement:

The data and code used in this study are currently confidential and not available for public access. We acknowledge the importance of data and code sharing in scientific research for validation and replication purposes. Therefore, we are considering options for future release. Our decision to maintain confidentiality at this stage is guided by ongoing research considerations and proprietary interests. We intend to review this decision in due course and will provide updates regarding the availability of our data and code as appropriate. Further inquiries about our data and code can be directed to the corresponding author.

Yuda Bi
Anees Abrol
Sihan Jia
Jing Sui
Vince Calhoun

Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS),
{ybi3, vcalhoun}@gsu.edu

Title page with author details

Gray Matters: ViT-GAN Framework for Identifying Schizophrenia Biomarkers Linking Structural MRI and Functional Connectivity

Yuda Bi*, Anees Abrol*, Sihan Jia*, Jing Sui*, Vince D. Calhoun*
*Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory, Atlanta, GA-30303, USA

Declaration of Interest Statement:

None of the authors have any conflicts of interest to declare in relation to this manuscript. This includes any financial, personal, or professional interests that could be construed to influence the work reported in this paper. We confirm that the content of the manuscript has not been influenced by any external interests, and all research was conducted in accordance with ethical standards. This statement is true to the best of our knowledge and belief, and any potential conflicts will be disclosed promptly should they arise in the future.

Yuda Bi
Anees Abrol
Sihan Jia
Jing Sui
Vince Calhoun

Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS),
{ybi3, vcalhoun}@gsu.edu