Received: 4 June 2023 | Revised: 18 August 2023 | Accepted: 22 August 2023

DOI: 10.1002/brx2.29

REVIEW ARTICLE

Brain-X Open Access WILEY

Understanding the brain with attention: A survey of transformers in brain sciences

Cheng Chen[1] | Huilin Wang[1] | Yunqing Chen[1] | Zihan Yin[2,3] | Xinye Yang[1] | Huansheng Ning[1] | Qian Zhang[2,3] | Weiguang Li[4] | Ruoxiu Xiao[1,5] | Jizong Zhao[2,3]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

[2]Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China

[3]China National Clinical Research Center for Neurological Diseases, Beijing, China

[4]Department of Health Technology and Informatics, Hong Kong Polytechnic University, Hong Kong SAR, China

[5]Shunde Innovation School, University of Science and Technology Beijing, Foshan, Guangdong, China

Correspondence

Ruoxiu Xiao and Jizong Zhao.
Email: xiaoruoxiu@ustb.edu.cn and zhaojizong@bjtth.org

Funding information

National Natural Science Foundation of China, Grant/Award Number: 62176268; Fundamental Research Funds for the Central Universities, Grant/Award Number: FRF-TP-22-047A1; China Postdoctoral Science Foundation, Grant/Award Number: 2023M730226; Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences, Grant/Award Number: 2020-JKCS-008; Major Science and Technology Project of Zhejiang Province Health Commission, Grant/Award Number: WKJ-ZJ-2112

Abstract

Owing to their superior capabilities and advanced achievements, Transformers have gradually attracted attention with regard to understanding complex brain processing mechanisms. This study aims to comprehensively review and discuss the applications of Transformers in brain sciences. First, we present a brief introduction of the critical architecture of Transformers. Then, we overview and analyze their most relevant applications in brain sciences, including brain disease diagnosis, brain age prediction, brain anomaly detection, semantic segmentation, multi-modal registration, functional Magnetic Resonance Imaging (fMRI) modeling, Electroencephalogram (EEG) processing, and multi-task collaboration. We organize the model details and open sources for reference and replication. In addition, we discuss the quantitative assessments, model complexity, and optimization of Transformers, which are topics of great concern in the field. Finally, we explore possible future challenges and opportunities, exploiting some concrete and recent cases to provoke discussion and innovation. We hope that this review will stimulate interest in further research on Transformers in the context of brain sciences.

KEYWORDS

brain science, EEG, fMRI, MRI, transformer

1 | INTRODUCTION

The brain is the most complex and mysterious organ in living things.[1,2] Brain sciences, which focus on understanding brain structure and function, have become an important research field of global interest.[3,4] By clarifying the brain's structural and functional characteristics, biological and psychological phenomena can be explained, allowing for the establishment of application-based fields such as human-computer hybrid interaction,[5] brain-like intelligence,[6] and the treatment of neurological diseases[7]; examples of progress in this regard include brain-computer interfaces,[8] cognitive neuroscience,[9] and the development of wearable equipment.[10] However, the modeling and analysis of brain characteristics remain challenging owing to the brain's complexity of function and structure.

Large-scale computational models are useful tools for explaining the complex computational mechanisms of the

brain. Recently, Transformers[11-13] have achieved significant success in text parsing,[14-16] object detection,[17,18] and visual segmentation.[19,20] Their excellent properties, including attention mechanism and extensibility, provide vital technical support for exploring complex brain mechanisms. As they have exhibited remarkable capabilities in cerebrovascular segmentation,[21] brain tumor segmentation,[22] Electroencephalogram (EEG) processing,[23] and brain age prediction,[24] Transformers are becoming important models for investigations in brain science.

Compared with Convolutional Neural Networks (CNNs),[25-27] Transformers abandon translation invariance and design their architecture using the self-attention mechanism. Although a part of their local expression ability is lost, this approach can better engender global context information for different domain spaces, which generates important attributes for brain sciences, such as the various kinds of connections between neurons,[28] EEG transmission process[29] and the basis of neural circuits[30] for brain function. In addition, dynamic adaptation for Transformer parameter attributes makes it possible to gain multivariate features by matching unlimited parameter space,[31,32] which makes Transformers highly adaptable to big datasets and can significantly improve their performance with increases in parameter magnitude, giving rise to hyper-heavy Transformers with over 100 million parameters.[33-35] Therefore, Transformers have the potential to build more complex models for the field of brain sciences in the near future.

As Figure 1 illustrates, this study aims to comprehensively review and summarize the most representative works regarding Transformers in brain sciences, focusing on modeling and calculation with Transformers in different areas. Depending on the scientific task and its technical solution, we divide the research on Transformers into different application-based scenarios and recorded important parameters to facilitate browsing and tracing by readers. In addition, we summarize and discuss the quantitative

Key points

What is already known about this topic?

- Transformers have promoted the rapid development of complex brain science tasks, and a detailed review thereof can be a valuable reference for the research field. Existing reviews on Transformers are arranged in terms of broad natural language or images and lack an in-depth discussion in the context of brain sciences. Therefore, a comprehensive review of Transformers in brain sciences is warranted and worthwhile.

What does this study add?

- To focus on the development of Transformers in brain sciences, this review comprehensively covers eight widely relevant application scenarios explored in brain sciences and discusses their quantitative assessments and technical details. By also describing the associated challenges and prospects, it aims to provide an overview of the latest research developments in brain science and is expected to contain a widely applicable reference and inspiration for researchers in related fields.

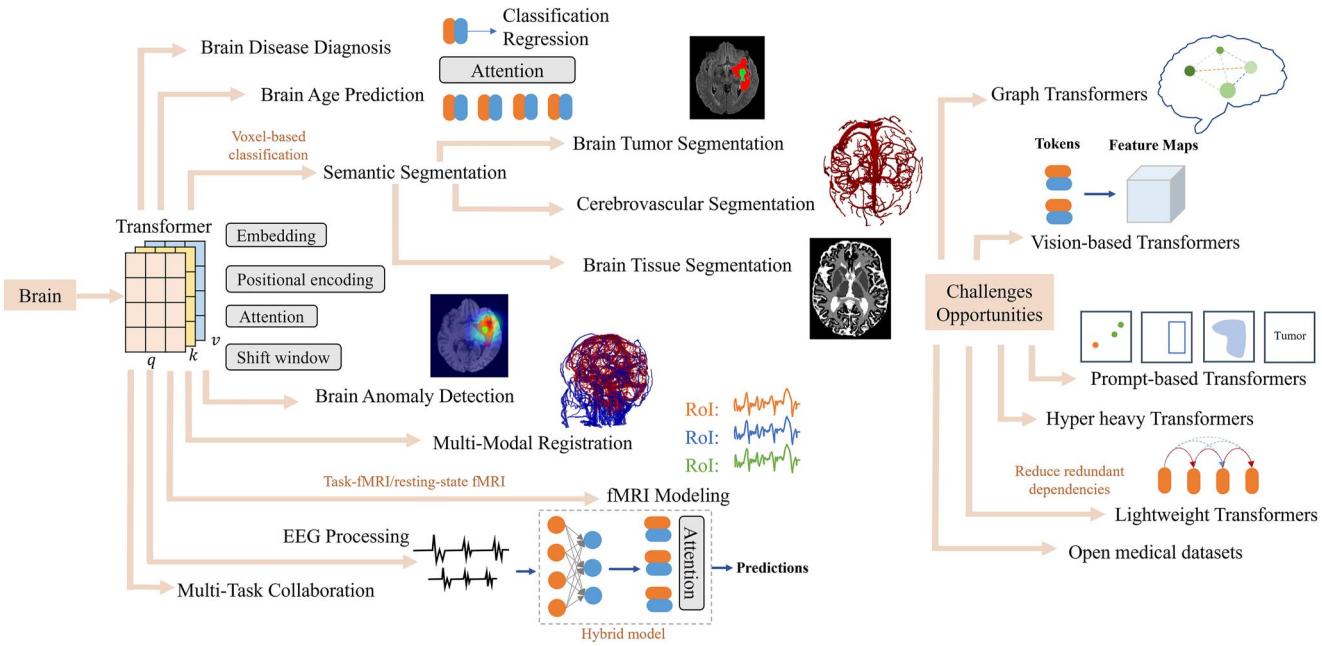

FIGURE 1 An illustration of Transformers in brain sciences. Gray blocks indicate important Transformer architectures, and orange arrows highlight diverse Transformer applications and opportunities.

assessments and model complexity of some focal challenges in brain sciences and prospect the possible challenges and opportunities. We hope that this review can serve as a valuable reference for researchers in related fields.

We organized this review as follows: Section 2 briefly introduces the basic architecture of Transformers. Section 3 reviews and analyzes the most relevant applications of Transformers in brain sciences, which are as follows:

- Brain disease diagnosis
- Brain age prediction
- Brain anomaly detection
- Semantic segmentation
- Multi-modal registration
- Functional Magnetic Resonance Imaging (fMRI) modeling
- EEG processing
- Multi-task collaboration

Section 4 discusses the quantitative assessments and model complexity of some representative computing tasks with Transformers. Section 5 puts forward the main challenges and potential opportunities with regard to Transformer use. The final section concludes this study.

2 | BASIC ARCHITECTURE OF TRANSFORMERS

Transformers are deep learning models based on the attention mechanism.[36,37] With continuous development and innovation, Transformers have gained many advanced structures; examples of such Transformers include Transformer in 3D CNN for 3D MRI Brain Tumor Segmentation (TransBTS),[38] UNEt TRansformers (UNETR),[19] Generative Pre-trained Transformer (GPT),[15,16,39] and Bidirectional Encoder Representations from Transformers (BERT).[14] Although they have different designs and interfaces for different tasks, they all have similar core structures. Figure 2 illustrates the core structure of the Transformer, and a brief introduction to the most common architectures in Transformers is given below.

2.1 | Input embedding

2.1.1 | Word embedding

Word embedding transforms the text sequence into vector representations for Transformers to capture the relationships between words.[40] One-hot encoding is a classic method used for categorical data; however, it is very sparse when faced with a large thesaurus.[41,42] To improve computing efficiency, word embedding exploited a learnable weight matrix to embed words into a numeric vector space. As the weights are updated, words with similar semantics gradually become close to each other in coding vectors.[43] Because of the property of the matrix transpose, word embedding also has the effect of feature dimensionality reduction.

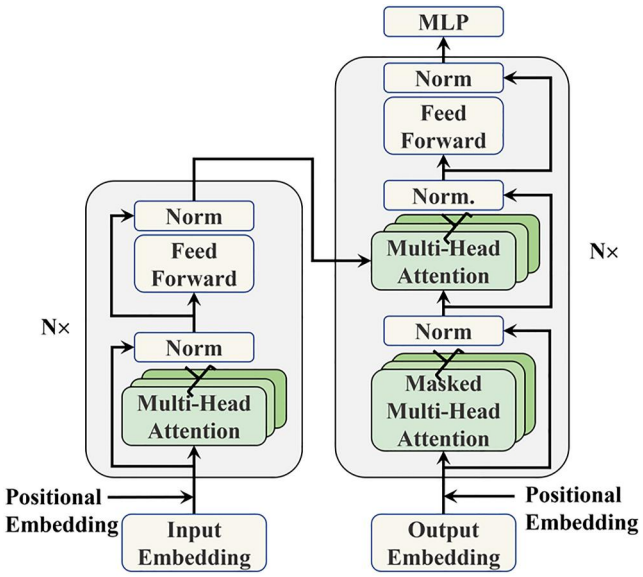

FIGURE 2 A transformer with stacked attention modules.

2.1.2 | Image embedding

Since standard Transformers only accept 1D vectors, images require dividing and flattening before their transfer to Transformers.[44] A common method of image embedding is patch embedding, which divides the images into 2D patches and then maps them onto the 1D channel.[45] Patch embedding is completed using a convolution operation whose kernel size and stride are consistent with the patch size. Similarly, 3D patches can be cropped and flattened to complete the embedding.[19,46] In addition to process a 3D patch, 3D volume can separate the slices to be flattened from different sampling angles,[47,48] which depends on whether Transformers pay more attention to the correlation between patches or that between slicers. In addition, patch merging is a calculation similar to a pooling operation and does not require any convolutions.[49] It averages or maximizes the window to compress features and concatenates them into vectors. Swin Transformers[49] introduced patch-merging embedding without any convolution operation, which groups patches and concatenates them along the depth to implement multiple sampling.

2.2 | Positional encoding

Compared with the sequential computation of a Recurrent Neural Network (RNN),[50,51] Transformers implement parallel computation using a matrix data stream, thus causing the loss of sequence order, which is required to endow spatial information with vectors using positional encoding.[52] Common positional encoding includes learnable matrices or typical functions; the trigonometric function is a classic example of position encoding.[53] Because each word matches sine and cosine curves of different periods using the transformation equation of the trigonometric function, different positions obtain unique positional encoding. In addition, the

latest research reports on advanced positional encoding, such as Decoupled posItional attEntion for Transformers (DIET)[54] and Position Encoding Generator (PEG).[55]

2.3 | Attention mechanism

The attention mechanism forms the core of Transformers. Compared to the limited windows in the RNN,[50] Gated Recurrent Unit (GRU),[56,57] and Long Short-Term Memory (LSTM),[58,59] the attention mechanism has a theoretically infinite window and computing space. In the attention module, the word vector passes through three fully connected layers to create the query (q), key (k), and value (v) vectors. Subsequently, the q and k vectors are multiplied by the dot product matrix to produce the score matrix, which provides instructions on how one word should be related to other words. Then, the matrix is scaled for a more stable gradient in case of the explosive effect of multiplication. The SoftMax function[60] activates the attention weights, where higher scores are enhanced and lower scores are suppressed. Finally, the q vectors are weighted using the attention weights to calculate the enhanced output vector.

2.4 | Multi-head attention

To transform multi-head attention, Transformers establish multiple, fully connected layers to calculate parallel q, k, and v vectors. Each group of q, k, and v vectors forms an attention enhancement, which is called a head of multi-head attention. Therefore, the parallel multi-head maps vectors into various subspaces. In theory, the diverse subspaces have opportunities to learn different features to provide more awareness for Transformers, which endows the Transformers with marked extensibility.[34,35] Residual connection and layer normalization[25] are often followed by multi-head attention for stabilization and optimization. The feed-forward layer composed of the linear layers and Rectified Linear Unit (ReLU) activation[61] further processes vectors for more attention expression.

2.5 | Masked multi-head attention

Masked multi-head attention is usually inserted in the decoders of Natural Language Processing (NLP) tasks.[62,63] As an autoregressive processor, the decoder takes a list of previous outputs and vectors from the encoder as input vectors. The decoding stops when the decoder generates an end flag.[64] To focus on the correlation of previous output vectors, the construction of multi-head attention is also required. Because of the autoregressive property, Transformers generate sequences word-by-word; hence, a look-ahead mask is necessary to prevent the conditional processing of future flags. This is a matrix of the same size as the attention weight, which leaves a zero-attention score for the future flags, driving the model to no longer focus on the future output vectors.

2.6 | Swin Transformer

The Swin Transformer[49] is the most typical variation of Transformers and has performed remarkably in many tasks owing to its multi-scale receptive field and efficient computing performance. Its core structure is as follows:

2.6.1 | Patch merging

The Swin Transformer[49] abandons fixed-size image embedding in the Vision Transformer (ViT)[45] and adopts patch merging to realize down-sampling. Patch merging takes the maximum or average value in the selected window, mapping the window region into the tensor in multiple channels, then adjusts the channels using the linear calculation to provide the word vector for Transformers. Therefore, patch merging can maintain the locality prior to CNN in the hierarchical network structure design to expand the receptive field of patch nodes.

2.6.2 | Window-based Multi-head Self-Attention (W-MSA)

Considering that multi-head attention may establish redundant dependencies over long distances, Swin Transformers[49] proposed a shifted window-based multi-head attention module. As illustrated in Figure 3A, the shifted windows split the feature maps into diverse attention partitioning. Transformers only calculate the attention weights within the partitioning. Therefore, shifted window-based multi-head attention focuses on hierarchical feature maps and has linear computational complexity.

2.6.3 | Shifted Window-based Multi-head Self-Attention (SW-MSA)

Although W-MSA effectively relieves computational pressure, it also leads to the loss of the correlation between windows. SW-MSA exploits the shifted windows to gain cross-window information and establishes a cyclic shift to correct inconsistent window scales (see Figure 3B). Therefore, the cyclic shift of the shifted windows applies more flexible modeling for various semantic scenarios and ensures a correlation between the shifted windows. SW-MSA further masks the weights in the cyclic shifted windows and calculates the attention enhancement.

3 | TRANSFORMERS IN BRAIN SCIENCES

3.1 | Brain disease diagnosis

Brain diseases usually have empirical manifestations seen on brain imaging, such as cerebral infarction, cortical atrophy,

multiple gray and white matter foci, vasculitis, and nodules.[65-67] Therefore, brain imaging has become the most important tool to detect and diagnose brain diseases. Magnetic Resonance Imaging (MRI) is a commonly used brain imaging method owing to its advantages of no radiation use, high resolution, and almost no allergic reaction.[68,69] MRI images for brain disease examination commonly contain T1-weighted (T1), T1-weighted imaging with Contrast Enhancement (T1-CE), T2-weighted (T2), and FLuid Attenuated Inversion Recovery imaging (FLAIR) volumes.[70] Since brain disease diagnosis depends on the professional knowledge and clinical experience of doctors, efficient auxiliary diagnostic methods play a critical role in clinical practice.

Deep learning is the most popular technique for brain disease diagnosis.[86,87] By modeling the association between brain images and disease, it fits a variety of complex mapping relationships and reduces dependence on feature extraction engineering, such as gray statistical histograms[88] and graph modeling maps.[89] This approach will show more robust and superior performance as datasets are extended. Since their development, Transformers have attracted wide attention in investigations of brain disease diagnosis, which creates stronger dependencies on multimodal data and the intrinsic relationships within the data. Dosovitskiy et al.'s[45] exploitation of the ViT represented the first time Transformers were applied to computer vision. As illustrated in Figure 4, ViT has also become an important baseline for

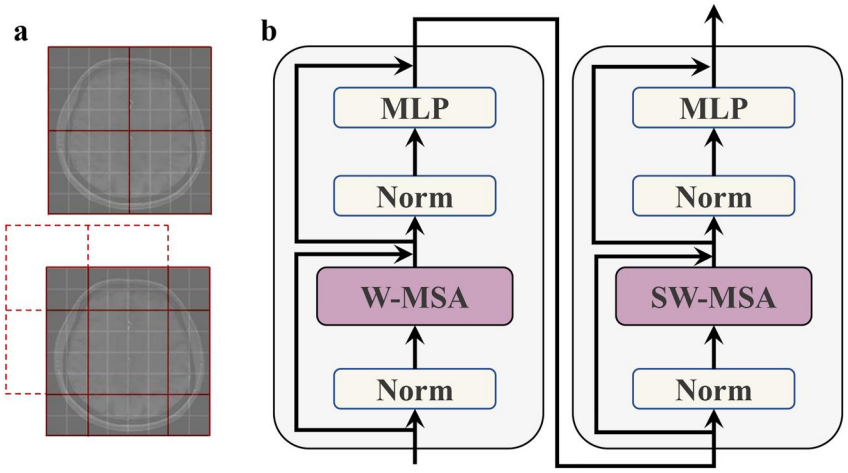

FIGURE 3 Core structure from Swin Transformer. (A) is the partition using shifted windows. (B) is the W-MSA and SW-MSA in the Swin Transformer.

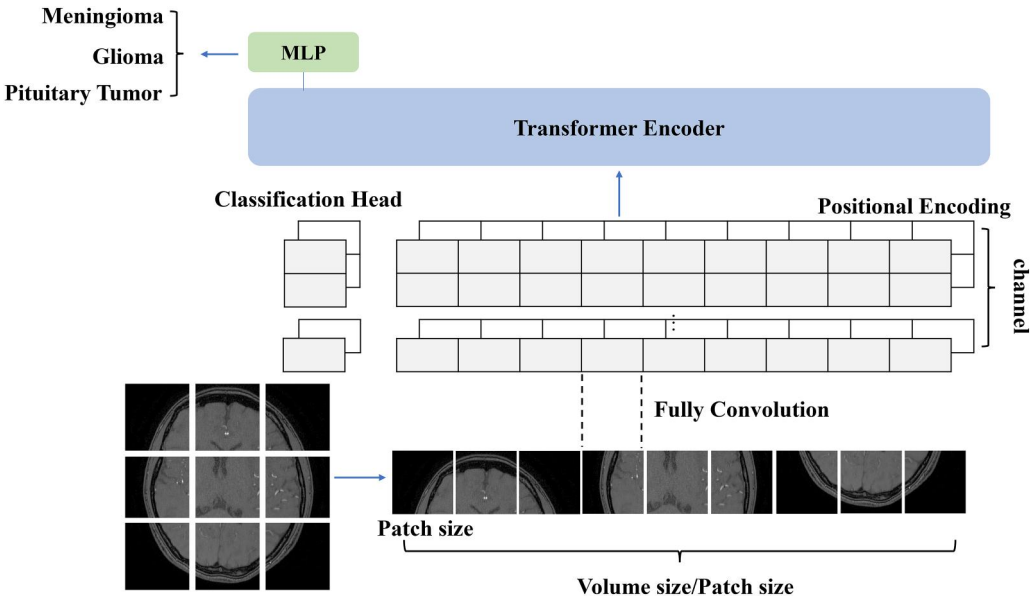

FIGURE 4 Brain disease diagnosis using the ViT. The brain image is cropped as diverse patches, which are converted to tokens using linear projection. In the Transformer encoder, the model captures attention for tokens to establish dependencies and output diagnostics.

brain disease diagnosis. For example, Odusami and Shin et al.[71,73] carried out the diagnosis of Alzheimer's Disease (AD) using a fine-tuned ViT. They focused on exploring the prediction of AD and amyloid changes via Positron Emission Tomography (PET), since AD biomarkers are an important reference for no or Mild Cognitive Impairment (MCI).[90] Odusami et al.[71] fused MRI with [18F]-FluoroDeoxyGlucose (FDG) PET[91,92] for richer semantics. Shin et al.[73] were the first to apply the ViT to the [18F]-FlorBetaBen (FBB) PET dataset.[93] Tummala et al.[74] exploited the fine-tuned ViT in the diagnosis of brain tumors, which offered support to radiologists for decision-making. The ViT retains the complete Transformer encoder, turning patches into tokens to build self-attention enhancement. An additional token is exploited to define a classification header, which establishes long-distance dependencies with image patches, outputting diagnostic decisions. This approach is conducive to module encapsulation and simple implementation and has become an important backbone of Transformers.

Inspired by the ViT, more Transformers with visual missions have been developed. Li et al.[77] developed a ViT-based model for brain tumor Whole-Slide Image (WSI) analysis (ViT-WSI). They collected patches by scanning the surgical specimens of patients and recorded molecular labels as predictions. Due to the abundance of label information for brain tumor type and subtype, a high degree of interpretability is provided for the ViT-WSI. Qiu et al.[78] proposed a Multi-channel Sparse Graph Transformer Network (MSGTN) that models static and dynamic information between individuals and groups, acquiring a rich feature representation. By obtaining a weighted fusion of multi-channel information, Transformers perform the clustering of each node to achieve early AD identification. Aloraini et al.[79] proposed a hybrid Transformer-Enhanced CNN (TECNN) for brain tumor classification. They built dual paths of a CNN and a Transformer for local and global guidance, improving long-distance dependency. Similar to,[79] Zhou et al.[80] proposed the Adaptive Sparse Interaction Residual Network (ResNet)-ViT Dual-Branch Network (ASI-DBNet) for the histopathological grading of brain cancer. The ASI-DBNet exploited parallel ResNet[25] and ViT to capture local and global information. The Adaptive Sparse Interaction Block (ASIB) with an embedded attention mechanism was designed to mediate better interaction of ViT and ResNet data streams and filter redundant features. Anaya-Isaza et al.[81] introduced a Cross-Transformer for brain tumor classification and tumor detection. They crossed vectors to calculate attention weights in different attention heads, flexibly learning more important pathological features. Considering that Transformers rely on large-scale datasets for self-attention calculation, Ferdous et al.[82] proposed a Linear-Complexity Data-Efficient image Transformer (LCDEiT), which used the teacher-student strategy and external attention mechanism to gain low-complexity calculation, leading to rapid brain tumor classification. Sarasua et al.[83] introduced a spatio-temporal network for 3D anatomical Meshes (TransforMesh) that exploited Transformers to incorporate heterogeneous trajectories, which led to the prediction of neuroanatomical changes. This was also the first time that Transformers were combined with mesh networks. Samak et al.[84] proposed a Transformer-based multimodal network (TranSOP) that incorporated attention enhancement for 3D Non-Contrast Computed Tomography (NCCT) imaging features. In addition, TranSOP incorporated clinical meta-information to enhance a priori modeling for the modified Rankin Scale (mRS) calculation of stroke outcomes. Zhou et al.[85] carried out tumor identification in Microsurgical Aneurysm Clipping Surgery (MACS). They developed vidMACSSwin-T for video frame classification to detect aneurysms, thereby enabling the surgical team to be aware of high-risk moments and take preventive action.

Table 1 presents these representative Transformers in brain disease diagnosis. The methods described above are often applied in the diagnosis of brain tumors and AD and have generally improved diagnostic efficiency through the remarkable performance of Transformers. In these models, the Transformer encoder often implements modular integration,[77,82,84] which not only retains the complete function to stabilize its performance but also obtains different parameter matrices according to specific tasks, improving the expressiveness of downstream tasks. Some dual-branch designs retain CNNs and Transformers to build local and global feature extraction,[79,80] making up for the lack of single feature extraction. We also note that Cross-Transformer computes attention across different heads to get cross-space information,[81] which will help build new attention mechanisms. Some Transformers also have the expansive property; for example, the rich intermediate computations from ViT-WSI[77] are available for downstream interpretation (feature discovery, attribution analysis), which is more beneficial for clinical decision making. In addition, TranSOP[84] requires a different activation and output interface for calculating continuous values, which also facilitates migration to other regression tasks. Therefore, Transformers are gradually taking on an important auxiliary role in brain disease diagnosis.

3.2 | Brain age prediction

Brain age prediction monitors the regularity of brain changes, contributing to the exploration of the correlation between brain characteristics and actual brain age in patients with brain diseases.[24] Over the past few years, deep learning has been successfully exploited to model the mapping between brain structure and brain age, demonstrating that brain structure (especially based on MRI) is associated with brain age.[94-96]

Due to their effective remote modeling of geometry and feature space, Transformers provide important clues to the correlation of brain images and have gradually become vital to researchers in this field. For example, He et al.[97] proposed a Global-Local Transformer for brain age prediction in

TABLE 1 Details and quantitative assessments of Transformers in brain disease diagnosis.

Author	Method	Year	Application	Modality	*N*	Loss	Optimizer	Acc. (%)	F1 (%)
Odusami et al.[71]	ViT[45]	2023	Early detection of AD	T1	100	-	AdamW[72]	93.75	-
				FDG	100				
Shin et al.[73]	ViT[45]	2023	Classification of AD	FBB	150	-	-	56.67	54.55
								80.00	66.67
Tummala et al.[74]	ViT[45]	2022	Tumor classification	T1	3064	CE	RMSprop[75]	98.21	-
							Adam		
							Adadelta[76]		
Li et al.[77]	ViT-WSI[77]	2023	Tumor classification	WSI	-	CE	Adam	-	-
Qiu et al.[78]	MSGTN[78]	2021	Identification of AD	fMRI/DTI/non-image (gender, age)	170	-	-	92.12	91.13
Aloraini et al.[79]	BECNN[79]	2023	Tumor classification	T1/T1-CE/T2/FLAIR	1425	-	-	96.75	96.80
					3064			99.10	98.70
Zhou et al.[80]	ASI-DBNet[80]	2023	Tumor progression	Microscopic image	2500	CE	AdamW	95.24	-
Anaya-Isaza et al.[81]	Cross-transformer[81]	2023	Tumor detection	T1	3929	CE	Aadelta	88.06	82.89
				FLAIR				89.58	84.76
				T1-CE				88.31	82.84
Ferdous et al.[82]	LCDEiT[82]	2023	Tumor classification	T1-CE	233	CE	AdamW	98.11	93.69
				T1/T1-CE/T2/FLAIR	2040			97.86	93.68
Sarasua et al.[83]	TransforMesh[83]	2021	Detection of neuroanatomical changes in AD	T1/T2*/FLAIR	-	MAE with regularization constant	-	-	-
Samak et al.[84]	TransSOP[84]	2023	mRS score	NCCT	500	-	Adam	80.00	59.00
				Clinical metadata					
Zhou et al.[85]	vidMACSSwin-T[85]	2023	Detection of cerebral aneurysms	Surgery video	356,165	Weighted CE	AdamW	87.10	58.90

Abbreviations: AD, Alzheimer's Disease; Adam, Adaptive moment estimation; AdamW, Adam Weight decay; CE, Cross Entropy; DTI, Diffusion Tensor Imaging; F1, F1-score; FBB, [18F]-FlorBetaBen; FDG, [18F]-FluoroDeoxyGlucose; FLAIR, FLuid Attenuated Inversion Recovery imaging; fMRI, functional Magnetic Resonance Imaging; MAE/L1, Mean Absolute Error; NCCT, Non-Contrast Computed Tomography; RMSprop, Root Mean Square propagation; T1, T1-weighted imaging; T1-CE, T1-weighted imaging with Contrast Enhancement; T2*, T2*-weighted imaging; T2, T2-weighted imaging; WSI, whole-slide images.

MRI. They designed a global pathway and local pathway to input complete MRI information and obtain local random patches, respectively. The Global-Local Transformer calculated attention in the global pathway and a weights map in the local-pathway, constructing local fine-grained details in the global context for comprehensive regression calculation. Cai et al.[98] proposed a graph transformer geometric learning framework for brain age prediction using Structural MRI (sMRI) and Diffusion Tensor Imaging (DTI). To deal with irregular brain Regions of Interests (RoIs), they transformed the feature maps into 90 RoI nodes via templates and aggregated them into feature vectors. The Transformer learned cross-modal interactions and fusion using geometric learning to achieve brain age estimation. Hu et al.[99] introduced the SQueeze and Excitation Transformer (SQET). They combined squeeze and excitation modules[100] in the self-attention mechanism to reduce computational complexity and added an Inception-based pyramid module[101] to exploit fine-grained local features and pursue accurate age prediction.

We summarize the main details of these brain age prediction methods in Table 2. The Global-Local Transformer exploits a dual-branch design to provide different parameter matrices for global and local awareness.[97] Considering the high computational cost of Transformers, the squeeze and excitation pretreatment[99] and the extraction of RoIs[98] significantly improve computational efficiency. We note that the T1 is the primary choice to explore the correlation between changes in anatomical structure and brain aging, as evidenced by studies on premature brain aging and cognition.[102,103] In

TABLE 2 Details and quantitative assessments of Transformers in brain age prediction.

Author	Method	Year	Modality	*N*	Loss	Optimizer	MAE	PCC (%)
He et al.[97]	Global-local transformer[97]	2021	T1	8379	MAE	Adam	2.70	98.53
Cai et al.[98]	Graph transformer[98]	2022	T1/DTI	16,458	MSE	Adam	2.71	86.82
Hu et al.[99]	SQET[99]	2022	T1	6318	MAE	AdamW	2.55	98.30

Abbreviations: Adam, Adaptive moment estimation; AdamW, Adam Weight decay; DTI, Diffusion Tensor Imaging; MAE/L1, Mean Absolute Error; MSE/L2, Mean Square Error; PCC, Person's Correlation Coefficient; SQET, SQueeze and Excitation Transformer; T1, T1-weighted imaging.

the regression loss function, the Mean Absolute Error (MAE) is a common choice because of its robustness and reduction of loss shock caused by abnormal points. Owing to its simplicity, Mean Squared Error (MSE) is prone to be adopted for large-scale samples.

3.3 | Brain anomaly detection

Brain anomaly detection research is active for accurate and quick region detection, assisting doctors in locating brain disease areas. We roughly divide brain anomaly detection into two categories: bounding box-based and reconstruction models.

The bounding box-based model is a classic method to describe the spatial location of abnormal objects.[104,105] Given a set of images, it analyzes the feature distribution to predict a set of object boxes. Its main benefits for Transformers include eliminating the dependency on candidate boxes such as the Region Proposal Network (RPN).[106] For example, Li et al.[107] proposed a View-Disentangled TransFormer (VD-Former) that collected the dense correlation of 3D brain positions to simulate contrast and spatial coherence. They deployed the VD-Former in a view-disentangled detection model to locate injury areas in brain MRI slices. Predictably, some advanced object detection models (such as DEtection TRansformer [DETR] and Performers[108]) will also demonstrate superior performance owing to their mature theoretical system.

The reconstruction model is another anomaly detection method with the advantage of weak supervision or no supervision. It often yields brain features in pre-training models that can be transferred to learn representation for normal samples[109-111] so that the anomaly detection of unknown pathological regions can be achieved in the reference process. Some statistical analysis methods are also implemented based on these pre-trained parameters.[112,113] For example, Chen et al.[114] introduced a U-TRansformer-based Anomaly Detection (UTRAD) framework that implanted a deep pre-trained model[25] to obtain dispersed vectors, integrating a Transformer to obtain stronger attention. They chose to learn the reconstruction representation in the feature distribution instead of raw images to obtain more features. Costa et al.[115] utilized the reconstruction model to gain the pattern distribution of the healthy brain. They built a Transformer to learn the deviations or outliers in pathological regions, thereby detecting subtle changes in the brain caused by early schizophrenia. Pinaya et al.[116] trained a Vector Quantized Variational AutoEncoder (VQ-VAE) to define the distribution of healthy brain data. Subsequently, they used a Transformer to model the compactness, complexity, and interaction of discrete elements to achieve unsupervised anomaly detection by defining healthy brain imaging data deviations.

Compared to disease diagnosis, anomaly detection requires determination of the presence of lesions and locating their approximate boundary. In the design of Transformers, the reconstruction errors of pre-trained parameters and enhanced feature maps need to be fully considered and the spatial information in the resamples must be constructed to perceive anomalies, a more challenging process. Transformers contribute to improving existing brain anomaly detection methods by introducing rich intrinsic correlations of brain imaging. Hence, it will be beneficial to promote further research on the characteristics of Transformers in the context of brain anomaly detection.

3.4 | Semantic segmentation

Semantic segmentation is a pixel-based classification of the target region that accurately describes the target structure.[117,118] It has become critical in brain imaging for downstream tasks, such as lesion measurement, 3D reconstruction, surgical navigation, and computational simulation.[119,120] It is fruitful for CNNs to establish long-term context information by means of continuous convolution.[121,122] Even with the deepening of the model structure and the abstract representation of multi-channel feature maps, CNNs show limitations in global feature modeling. Compared to CNNs, Transformers are more appropriate for global dependency modeling, providing vital clues for complex visual understanding, especially for intensive segmentation tasks. The semantic segmentation of brain imaging mainly includes brain tumor segmentation, cerebrovascular segmentation, and brain tissue segmentation.

3.4.1 | Brain tumor segmentation

The U-shaped structure is the backbone of brain tumor segmentation, having the advantages of symmetry and high resolution.[123,124] A typical U-shaped Transformer for brain images is UNETR,[19] which embedded 3D volume into a

Transformer using 3D patch compression. Its decoder incorporated the CNN structure to capture global multi-scale information and calculated the final semantic segmentation.[19] Hatamizadeh et al.[125] proposed a Swin UNEt TRansformer (Swin UNETR) model replacing the encoder with the Swin Transformer[49] to capture multi-scale features in the partitioning scheme with shifted windows. Similar to the Swin UNETR,[125] Liang et al.[126] embedded the Swin Transformer[49] into the encoder structure using parallel shift windows. They further supplemented prior knowledge to realize semantic modeling more efficiently. Inspired by Inception with parallel branch architecture,[127] Liang et al.[128] introduced a parallel module named Transformer-Convolution Inception (TC-Inception)-based UNETR to extract global features and thus expand the data stream for multi-scale semantic awareness. These models typically replace encoders with Transformers for more compact semantics, with stronger global context information than traditional CNNs.

Wang et al.[38] exploited the classic TransBTS. Its attention layer was at the bottleneck layer of CNN, digging the internal correlation in the deep potential space. As a hybrid of a CNN and a Transformer, TransBTS first exploited the 3D CNN to capture compact feature maps of space and channel and then used the Transformer to model long-distance dependencies in global space. Through the feature fusion and analysis from the Transformer and CNN, TransBTS produced high-resolution segmentation. Subsequently, an improved version of TransBTS called TransBTSV2 was developed,[129] which contained a variable bottleneck module with the Transformer for shape and detail awareness, thereby achieving superior performance. Lyu et al.[130] also replaced the bottleneck layer with 12 consecutive Transformers to obtain attention enhancement. Using TransBTS[38] as the main feature and core backbone, Jia et al.[131] proposed a CNN-Transformer combined model (BiTr-Unet), which was designed with a deeper structure and an embedded attention module in two successive layers at its deepest to enhance global dependency. Transformers derived from TransBTS are built at the bottleneck of the U-shaped structure. Since the self-attention calculation is completed at the bottleneck, the model retains the advantage of CNN and expands the global dependence of the deep feature maps.

Some advanced U-shaped Transformers are available for brain tumor segmentation. For example, Liang et al.[132] proposed a Swin Transformer-based network for Brain Tumor Segmentation (BTSwin-Unet) by embedding a Swin Transformer in each layer, which could enhance the characteristics in each calculation stage. Similarly, Jiang et al.[133] proposed a method for Brain Tumor Segmentation using Swin Transformer (SwinBTS). They also exploited the Swin Transformer to replace different layers in the encoder and decoder. In addition, they inserted an enhanced Transformer block to supplement the details at the bottleneck. Peiris et al.[134] proposed a Volumetric Transformer architecture (VT-UNet) for brain tumor segmentation. The encoder of VT-UNet calculated both local and global attention, while its decoder adopted parallel self-attention and cross-attention to capture and optimize boundary details. Li et al.[135] embedded a Transformer into UNet++[136] and used a dense feature extractor to model global feature information and remote dependencies. The U-shaped structure retains the advantage of a symmetry model, enabling the feature maps to establish a one-to-one matching relationship from shallow to deep layers. Transformers make semantic features denser to output high-resolution results.

In addition, some researchers abandoned U-shaped structures and built new Transformer architectures for more specific objectives. For example, Xing et al.[137] proposed a Nested modality-aware TransFormer (NestedFormer), which focused on intra- and inter-modal relationships in multi-modal MRIs. NestedFormer adopted a multi-branch structure to extract features from different MRIs and then completed fusion in Transformer-based feature aggregation for more effective representation. Zhang et al.[138] proposed a multi-modal medical TransFormer (mmFormer) that combined four hybrid modality-specific encoders for modal local and global context modeling and one modality-specific encoder for localizing modal invariant features and long-term correlations. Its convolutional decoder performed fusion and predicted tumor segmentation. Lin et al.[139] proposed a Clinical Knowledge-Driven Brain Tumor Segmentation model (CKD-TransBTS). They divided MRI into two groups based on imaging principles and designed a two-branch hybrid encoder to accurately extract the local features of the lesion boundary and remote features of 3D volume. Subsequently, they proposed calibration modules for the Transformer and CNN to reduce feature differences. Gai et al.[140] proposed a Residual Mix Transformer Fusion model (RMTF-Net). They added overlapping patch-embedded modules in the Transformer to enhance boundary coding capability. In addition, a parallel fusion strategy was proposed to obtain the encoding of local-global balance features in the Transformer. Zhu et al.[141] focused on semantic and edge information fusion in brain tumors. They constructed semantic segmentation and edge detection modules to extract deep semantics and edge features, respectively, which were constrained by the specific objective function. Finally, the feature fusion module combined different features to supervise the training for the Transformer. Nian et al.[142] developed a plug-and-play Infinite Deformable Fusion Transformer Module (IDFTM) that combined different attention heads using a fusion self-attention mechanism with attentional logic and weight mapping. The IDFTM can enhance spatial dependence in 3D volume and promote more advanced models. These models can be modeled more flexibly to fit the functions of Transformers in specific structures. The details and quantitative assessments of Transformers in brain tumor segmentation are presented in Table 3.

3.4.2 | Cerebrovascular segmentation

Cerebrovascular diseases continue to occur frequently and have become the main cause of death in human beings.[144,145]

TABLE 3 Details and quantitative assessments of Transformers in brain tumor segmentation.

Author	Method	Year	Data	Modality	*N*	Loss	Optimizer	Dice (%)	95HD (mm)
Wang and Li[38]	TransBTS	2021	BraTS 2019	T1/T1-CE/T2/FLAIR	460	Dice	Adam	83.62	5.14
			BraTS 2020		660	L2		83.52	10.89
Jiang et al.[133]	SwinBTS	2022	BraTS 2019	T1/T1-CE/T2/FLAIR	335	Dice	Adam	81.15	-
			BraTS 2020		494	CE		82.24	17.06
			BraTS 2021		1470			86.60	11.39
Xing et al.[137]	NestedFormer	2022	BraTS 2020	T1/T1-CE/T2/FLAIR	369	Dice	AdamW	86.10	5.05
			MeniSeg	T1GD/CE-FLAIR	110	CE		76.50	4.41
Hatamizadeh et al.[125]	Swin UNETR	2021	BraTS 2021	T1/T1-CE/T2/FLAIR	1470	Dice	-	88.97	5.21
Jia et al.[131]	BiTr-Unet	2021	BraTS 2021	T1/T1-CE/T2/FLAIR	1470	-	Adam	95.32	3.90
Liang et al.[128]	TransConver	2022	BraTS 2018	T1/T1-CE/T2/FLAIR	351	Dice	Adam	86.32	4.23
			BraTS 2019		460	CE		83.72	4.74
Zhang et al.[138]	mmFormer	2022	BraTS 2018	T1/T1-CE/T2/FLAIR	285	Dice	Adam	71.92	-
						CE			
Li et al.[129]	TransBTSV2	2022	BraTS 2019	T1/T1-CE/T2/FLAIR	460	Dice	Adam	85.18	4.87
			BraTS 2020		660	L2		84.90	7.45
Hatamizadeh et al.[19]	UNETR	2022	MSD BraTS	T1/T1-CE/T2/FLAIR	484	Dice	AdamW	71.10	8.82
						CE			
Peiris et al.[134]	VT-UNet	2022	MSD BraTS	T1/T1-CE/T2/FLAIR	484	-	AdamW	87.10	3.43
Lin et al.[139]	CKD-TransBTS	2023	BraTS 2021	T1/T1-CE/T2/FLAIR	1251	Dice	-	90.66	6.22
Gai et al.[140]	RMTF-Net	2022	LGG	T1/T1-CE/T2/FLAIR	1311	Dice	Adam	93.50	-
			BraTS 2019		10,047	CE		82.10	
			BraTS 2020		10,945	SSIM[143]		81.80	
Zhu et al.[141]	Z. Zhu	2023	BraTS 2018	T1/T1-CE/T2/FLAIR	285	Dice	Adam	86.93	4.19
			BraTS 2019		335	CE		88.22	4.02
			BraTS 2020		369			87.95	4.59
Liang et al.[132]	BTSwin-Unet	2022	BraTS 2018	T1/T1-CE/T2/FLAIR	351	MAE	Adam	86.40	4.37
			BraTS 2019		460	Dice		83.46	5.11
						CE			
Li et al.[135]	DenseTrans	2023	BraTS 2020	T1/T1-CE/T2/FLAIR	2000	-	SGD	86.00	12.81
			BraTS 2021		494		Adam	89.00	10.53
							MSG		
Nian et al.[142]	3D brainformer	2023	BraTS 2017	T1/T1-CE/T2/FLAIR	331	Dice	Adam	81.30	5.77
			BraTS 2018		351	L2		83.27	5.94
Liang et al.[126]	3D PSwinBTS	2023	MSD BraTS	T1/T1-CE/T2/FLAIR	484	Dice	Adam	57.80	-
			BraTS 2020		498	CE		84.94	10.75
			BraTS 2021		1470			87.32	10.78
Lyu et al.[130]	Q. Lyu	2022	Q. Lyu	T1/FSPGR	138	Dice	Adam	87.80	-
					145	CE		89.50	

Abbreviations: 95HD, 95% Hausdorff Distance; Adam, adaptive moment estimation; AdamW, Adam Weight decay; CE, Cross Entropy; CE-FLAIR, contrast-enhanced T2-FLAIR; Dice, Dice similarity coefficient; FLAIR, FLuid Attenuated Inversion Recovery imaging; FSPGR, Fast SPoiled GRadient echo; LGG, Low-Grade Glioma; MAE/L1, Mean Absolute Error; MSD, Medical Segmentation Decathlon; MSG, Momentum Stochastic Gradient; SGD, Stochastic Gradient Descent; SSIM, Structural SIMilarity; T1, T1-weighted imaging; T1-CE, T1-weighted imaging with Contrast Enhancement; T2, T2-weighted imaging.

Cerebrovascular segmentation is an important prerequisite for the measurement and navigation of cerebrovascular diseases.[21,146] In Transformer-based models, UNETR is an important baseline for cerebrovascular segmentation. For example, Chen et al.[147] and Wu et al.[148] successfully performed cerebrovascular segmentation using UNETR in time-of-flight Magnetic Resonance (MR) angiography (TOF-MRA). Chen et al.[147] further proposed a TRansformer with Semantic Fusion (TRSF-Net) for cerebrovascular segmentation. As shown in Figure 5, TRSF-Net modeled MRI as graph to enhance spatial dependence. The authors proposed semantic fusion to strengthen the correlation between the Transformer and the CNN, thereby realizing more efficient cerebrovascular segmentation. Li et al.[149] proposed a Vascular Segmentation method based on Swin Transformer (SegVesseler) to process whole-mouse-brain vascular datasets.[150,151] This approach established multi-resolution feature extraction to capture vascular texture at different scales, thus maintaining vascular connectivity and topology. Due to the characteristic of small structures, the modeling of cerebrovascular segmentation is still mainly based on the symmetry model, which relies on compact feature maps to obtain high-resolution perception and more detailed structure in the enhancement of Transformers. The details of the three abovementioned Transformers are presented in Table 4.

3.4.3 | Brain tissue segmentation

Following the operations from TransBTS, Rao et al.[152] proposed a Transformer-based Automated Brain tissue Segmentation (TABS) method to verify universality and reliability in four datasets. Zhang et al.[153] introduced a Stepped-Fusion Segmentation TransFormer (SF-SegFormer) framework that adopted semantic branches of pairwise structure to extract multi-modal features and further implemented multi-level fusion from cross-modal features to enrich the diversity of channel aggregation. Sun et al.[154] proposed a hybrid architecture with a CNN and a Transformer (HybridCTrm), which was designed as a multipath composed of convolution and a Transformer to encode input combination features, effectively combining and making full use of features from different modalities. Inspired by UNETR, Yu et al.[155] proposed a U-shaped model with Nested Transformers (UNesT). UNesT had a fast and simplified encoder design that enabled local communication between adjacent patches using hierarchical aggregation, thereby preserving relative location information. Zhang et al.[156] proposed a Transformer-Weighted Network (TW-Net), which adopted an encoder and decoder architecture. TW-Net placed the Transformer in the bottleneck layer, along with the introduction of squeeze-and-excitation blocks to integrate the boundary information of the brain structure. Additionally, the authors designed a hybrid loss of First Order Gradient (FOG),[157] topological,[158] and Cross-Entropy (CE) loss to improve optimization. The details and quantitative assessments of Transformers in brain tissue segmentation are given in Table 5. Transformers are mainly present in combination with CNNs in the segmentation of brain tissue (such as HybridCTrm[154] of dual-branch design and TABS[152] of bottleneck improvement) to obtain a better field of perception. Owing to their multi-modal form, Transformers introduce inter-class fusion between the multi-modes to achieve efficient feature representation. The

TABLE 4 Details and quantitative assessments of Transformers in cerebrovascular segmentation.

Author	Year	*N*	Loss	Optimizer	Dice (%)
Chen et al.[147]	2023	95	CE	Adam	78.83
Wu et al.[148]	2022	109	Hessian soft loss[148]	Adam	83.10
Li et al.[149]	2023	82	Dice	Adam	87.20
		100			96.23
		100			91.34

Abbreviations: Adam, adaptive moment estimation; CE, Cross Entropy; Dice, Dice similarity coefficient.

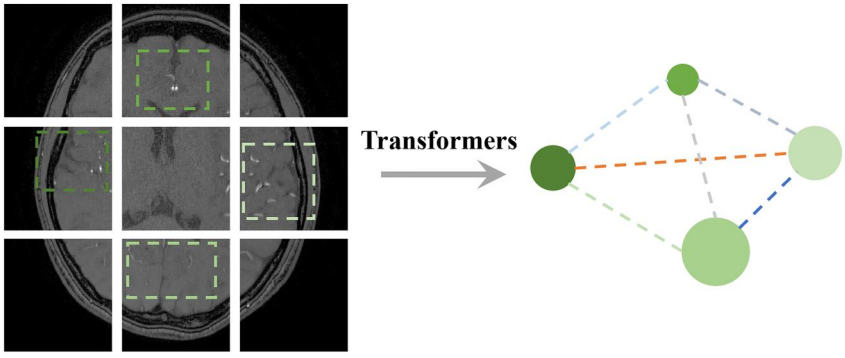

FIGURE 5 Based on graph theory, Magnetic Resonance Imaging patches are modeled as nodes and the Transformer calculates edges between nodes. The graph reflects the degree of correlation between different areas.

representative works discussed above prove that Transformers have promoted rapid innovation in brain semantic segmentation for rapid innovation. It is expected that further research will lead to the emergence of more and better Transformer-based brain semantic models.

3.5 | Multi-model registration

As Figure 6 illustrates, image registration involves the alignment of a moving image to a fixed image, which is conducive to multi-modal analysis and fusion, and has significance in the acquisition of multi-modal brain MRIs. Traditional image registration requires the establishment of effective feature detection and complex matching, and further generates transformation models to obtain deformation parameters; thus, it lacks a robust representation of image understanding and relies heavily on a priori theory.[159] Deep learning-based registration involves the construction of the global function via model learning, obtaining a representation of image registration alignment. Compared with the early demand to rely on the deformation field as an intermediate parameter,[160,161] unsupervised methods now receive more attention.[162,163] Due to the limitation of the receptive field, CNNs may obtain low-resolution features in deep layers, resulting in a lack of detailed positional information.[164,165] The long-range spatial relationship of Transformers can more accurately understand the spatial correspondence between the moving images and fixed images, making them strong candidates for image registration.

Recent research on registration is mainly divided into displacement field registration and diffeomorphic registration. To guide deformation, displacement field registration models the deformation field via deep learning to complete image transformation. For example, Chen et al.[166] used the Swin Transformer[49] to build the affine transformation network and generate transformation parameters that align the moving image to the fixed image. Furthermore, they introduced a hybrid Transformer-ConvNet model (TransMorph) to establish remote space correspondence between image pairs, thus realizing efficient deformable registration. Xu et al.[47] obtained a slice-to-volume reconstruction of fetal brain MRI. They built a volume registration method based on Transformers to model multiple stacks of MRIs as sequences. Then, they used attention to detect correlations between slices to predict slice transformations and alternately update volume. Chen et al.[167] proposed a hybrid CNN-Transformer architecture (ViT-V-Net) that integrated an affine transformation network and deformable registration network. They utilized an attention module to build up pixel-matching relationships between image pairs in latent spaces, thereby guiding the generation of deformed images. Lee

TABLE 5 Details and quantitative assessments of Transformers in brain tissue segmentation.

Author	Year	*N*	Loss	Optimizer	Dice (%)
Rao et al.[152]	2022	215	MSE	Adam	95.00
		283			95.27
		228			95.40
Zhang et al.[153]	2022	-	CE	Adam	99.52
Sun et al.[154]	2021	5	CE	-	83.47
		10			87.16
Yu et al.[155]	2022	263	-	-	74.44
		263			70.25
Zhang et al.[156]	2023	>9000	FOG[157]	-	86.00
		1440	Topological[158]		88.67
			CE		

Abbreviations: Adam, adaptive moment estimation; CE, Cross Entropy; Dice, Dice similarity coefficient; EEG, electroencephalogram; FOG, First Order Gradient; MSE, Mean Square Error.

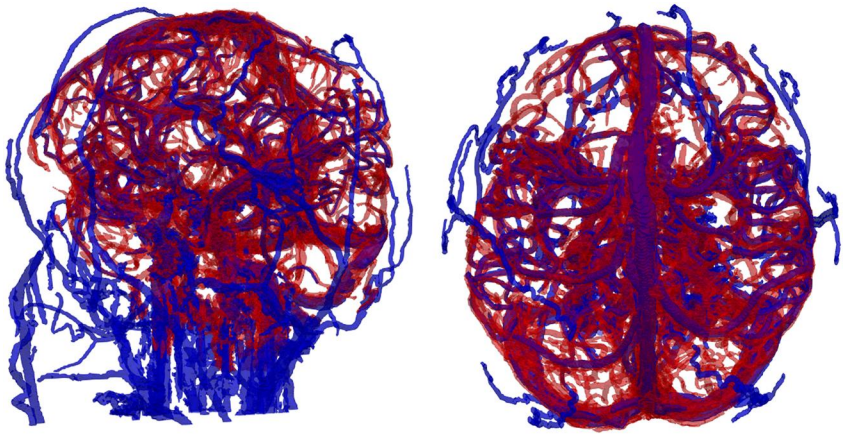

FIGURE 6 Cerebrovascular segmentation from phase-contrast MRA (PC-MRA; blue) and contrast-enhanced MRA (CE-MRA; red) can be correctly overlapped after registration.

et al.[168] introduced Image-and-Spatial Transformer Networks (ISTNs) toward Structures-of-Interest (SoI), learning the representation of deformed images with the proposed Transformer in a particular SoI to achieve accurate registration with limited training data. Mok et al.[169] proposed a Coarse-to-Fine Vision Transformer (C2FViT). They used a convolutional vision converter to achieve global and local connectivity and a multi-resolution strategy to obtain global affine, achieving fast and robust registration learning. Yang et al.[170] proposed a Graph convolution Transformer-based Deformable Image Registration (GraformerDIR) method for brain data. Considering that redundant and chaotic connections affect the effectiveness of remote dependencies, they constructed a graph neural network (GNN)[171] to establish graph nodes to capture local information, which aggregated features to obtain global information, thus obtaining the implicit association between nodes using the attention mechanism. Considering the high magnitude of parameters and computational complexity of Transformers, Ma et al.[172] proposed a Symmetric Transformer-based model (SymTrans) focusing on capturing local spatial context information to reduce semantic ambiguity, simulating remote spatial image pair correlation. Compared with CNN, Transformers better explain the spatial correspondence between fixed and moving images. Some Transformers embedded in CNNs produce a larger effective receptive field,[166] and the RoIs are freely and flexibly modeled to establish the feature-matching relationship.[170]

Unlike displacement field registration, the diffeomorphic method constructs a differential isomorphic deformation field to realize differential isomorphic properties. For example, Zhang et al.[173] proposed a Dual Transformer Network (DTN) for diffeomorphic registration that consisted of two Transformers to handle global correlation learning within and across images and finally generated deformed images using a registration reference module. Zhu et al.[174] proposed a symmetric unsupervised learning network based on the Swin Transformer (Swin-VoxelMorph) to minimize image dissimilarity and estimate transformations. In their Transformer, the hierarchical Swin Transformer with shifted windows extracted context characteristics, and a symmetric Swin Transformer estimated the registration fields with diffeomorphic properties, which were constrained by the objective function. In addition, Chen et al.[166] and Ma et al.[172] presented diffeomorphic variants for their Transformers, ensuring that the topologies retained deformation. These attention modules learn inter- and intra-image correlations and promote the semantic correspondence of anatomical structures in variational inference, thus obtaining a smooth and topology-preserving deformation.

Compared with other organs, the brain undergoes relatively small deformations due to the constraints of the skull, leading brain images to be the most relevant for image registration in the long term. Transformers pay attention to the relationship between images and long-range space, which is not restricted by the receptive field, thereby functioning as the backbone of image registration.

3.6 | fMRI modeling

fMRI (commonly task-fMRI) comprises T2-weighted imaging that is Blood Oxygenation Level-Dependent (BOLD) during certain brain activities. It is often applied to examine areas of functional localization such as movement, language, and sensation. Exploring the changing relationships between fMRI timing sequences is required to identify the target functional region.

Existing fMRI processing methods typically learn short sequence representations, allowing only short-range dependencies to be encoded.[175,176] However, more attention is being paid to remote-dependent modeling in brain informative law and specific motion recognition (e.g., motor imagery). Some of the recent breakthroughs in using Transformers to address the long-distance dependence of fMRI are discussed below. For example, Asadi et al.[177] proposed a Transformer to learn temporal and spatial context information in fMRI. They combined BOLD time series and functional connectivity networks into a Transformer to create dynamic dependencies that explicitly corresponded to brain functional RoIs. Malkiel et al.[178] introduced a Transformer framework (TTF) for the analysis of fMRI. TTF achieved 3D volume reconstruction using self-supervised training and was fine-tuned on specific tasks to achieve age and gender prediction and schizophrenia identification. Deng et al.[179] developed a Spatial-Temporal Transformer (ST-Transformer). After data balancing, ST-Transformer calculated the spatial and temporal representations in fMRI using a linear spatial-temporal multi-headed attention unit for the diagnosis of Autism Spectrum Disorder (ASD). Nguyen et al.[180] proposed a Brain Attend and Decode (BAnD) architecture to handle the task state decoding of fMRI. It exploited residual connections for spatial feature extraction. The self-attention mechanism strengthened feature correlation, realizing time dimension modeling. BAnD achieved conditional transfer using finetuning to accomplish specific tasks. Dai et al.[181] proposed an end-to-end functional interaction learning method (BrainFormer) that performed 3D convolution modeling for local cues and captured shallow to deep global attention, achieving a comprehensive diagnosis of brain diseases, including AD, depression, attention deficit hyperactivity disorder, and headache disorders. Bedel et al.[182] presented a Blood-oxygen-level-dependent Transformer model (BolT) for analyzing fMRI. BolT introduced a fused window attention mechanism to capture the edge attention of adjacent windows. It gradually increased the overlap range, transferring the local representation to the global representation. Li et al.[183] applied ViT to implement fMRI modeling and data analysis and demonstrated the importance of positional encoding. They proved that the Transformer contributed to the early diagnosis of ASD, AD, depression, and other neurological diseases. Yu et al.[184] proposed a twin-Transformer framework that built pairwise Transformers to analyze spatial and temporal information. The twin-Transformer framework deduced the functional network

using self-supervised training, and was successfully applied in motor task. To encode the regularity and variability of the brain, Zhao et al.[185] proposed a Temporal-Correlated AutoEncoder (TCAE) based on Transformers. They implemented learned embedding in the self-supervised learning to improve generalization and effectiveness for downstream task brain state prediction. Kumar et al.[186] utilized Transformers to simulate brain contextual meaning. They collected information from participants while they listened to semantic stories using BERT-base[14] and different semantic features to obtain fMRI sequences. Then, the BERT-base[14] and GPT-2[16] were built to update the vectors. The authors proved that the ability of the head to understand language corresponded with the brain activity in specific cortical areas.

Resting-state functional MRI (rs-fMRI) is a tool for functional brain imaging while the brain is in a resting, relaxed, and awake state and is applied to explore essential brain states. In investigating rs-fMRI, Hu et al.[187] focused on pre-training for Transformers and proposed a Transformer-based neural network (BrainNPT) that implemented unsupervised learning using replaced and masked RoIs. These self-supervised parameters were transformed to handle a variety of downstream tasks. Sarraf et al.[188] proposed an Optimized Vision Transformer for AD prediction (OViTAD). They collected rs-fMRI and sMRI to explore the function and structure of the brain in early-stage AD. Kan et al.[189] built vectors of a fixed size and order for modeling. Thus, their Transformer encoders focused on learning the strength of pairwise connections between RoIs at different layers. To account for similar behaviors, they compressed enhanced node characteristics into graph-level embeddings to capture the biological sex, brain disease, or other brain properties. These works have shown that rs-fMRI provides important evidence for mining the resting characteristics of the brain. Owing to its inactive target functional area, rs-fMRI is more commonly applied in reprocessing or guiding for structural images.

All the above representative studies have achieved significant success. The modeling of Transformers in fMRI is usually reflected in time series, using self-attention screening and enhancement to guide the complete process of modeling and analysis of fMRI to obtain target predictions. The studies also prove that Transformers contribute to exploring the feature relationships for different time series, which can more accurately identify target functional areas and provide a solution for the deep analysis of fMRI.

3.7 | EEG processing

An EEG signal is a pattern reflecting brain activity using electrophysiological indicators. It is usually collected by placing electrodes on the scalp. In some cases, deep electrodes are also required to obtain EEG signals of intracranial activity, such as in patients with intractable epilepsy.[190] The efficient and exact processing of EEG signals has attracted attention. Typically, the RNN-based model is used in traditional sequential signal processing.[191,192] However, it lacks long-term memory and parallel computing capability required to meet the current demand for more complex and explicable feature exploration, prompting researchers to develop Transformers for recognizing time series-based EEG signals.

Since EEG signals have a data structure similar to that of word sequences, researchers usually directly applied EEG signals for spatial-temporal correlation. For example, Song et al.[195] proposed a Spatial-Temporal Tiny Transformer (S3T) to focus on temporal transforming. Attention enhanced tokens were transformed into classification by global average pooling and fully connected layers, obtaining a highly distinguishable representation. Liu et al.[198] constructed four variant transformer frameworks with spatial, temporal, sequential spatial-temporal, and simultaneous spatial-temporal attention. Studies have shown that simultaneous spatial-temporal attention leads to the best emotion recognition performance in modeling EEG signals. Owing to its ease of collection and security, Du et al.[199] used EEG signals for personal identification. For this purpose, they encoded the temporal and spatial features and applied the attention mechanism to construct the coupling relationship. Wang et al.[200] proposed a Hierarchical Spatial Learning Transformer (HSLT) to capture brain spatial dependencies. They divided EEG signals into different brain regions according to Power Spectral Density (PSD)[206] and realized the hierarchical learning of spatial information from the electrode level to the brain region level using parallel self-attention. Tao et al.[202] established a Gated Recurrent Unit Transformer (GRUGate Transformer) to acquire the long-term dependence of EEG signals. The gating mechanism stabilized the GRUGate training process and was assessed in EEG datasets for human brain-visual and motor imagery. Siddhad et al.[203] implemented age and gender classification using a Transformer. The model could directly process EEG signals and discarded some redundant weights using Dropout.[207] Ahn et al.[204] proposed a multiscale convolutional Transformer that included spatial, spectral, and temporal domains for multi-head attention enhancement. They explained the experimental paradigm of motor imagery, visual imagery, and speech imagery tasks using different modalities.

In addition, some researches mapped EEG signals into feature maps using CNN architecture, implementing self-attention enhancement for latent features. For example, Wan et al.[193] proposed a Transformer-based EEG analysis model (EEGformer) for brain activity classification. They deeply convolved EEG signals to obtain channel features and calculated vectors from feature maps along the temporal, regional, and synchronous dimensions. Therefore, EEGformer could learn more about EEG interrelationships. Xie and Sun et al.[194,208] constructed hybrid model of a Transformer and CNN to describe the significance of temporal and spatial correlation in EEG signals. They verified that positional encoding effectively improved classification accuracy.

Moreover, they revealed that attention weight was consistent with the spectral analysis of mu and beta rhythms over the sensorimotor.[209,210] Lee et al.[196] applied a CNN to split feature tokens from EEG and added self-attention enhancement, which revealed the spatio-temporal characteristics of overt and imagined speech in EEG signals. They further adopted the squared Hinge loss[197] for robust classification. Kostas et al.[201] developed BERT-inspired Neural Data Representations (BENDR) for large-scale EEG data learning. BENDR developed brain-map modeling using self-supervised learning with a self-supervised loss-based cosine similarity and mean squared activation,[211] and was then fine-tuned to accommodate different downstream tasks. Similar to Ahn et al.,[204] Ma et al.[205] established a Transformer to focus on the global dependence of EEG in three domains. They placed a CNN in the post-processing of a spectral attention map to capture regional features and then implemented temporal attention enhancement to calculate global dependencies.

Table 6 records the critical details of diverse Transformers in EEG signal processing. Extensive research shows that Transformers have driven the rapid development of EEG processing owing to their sequence samplers. In time modeling, Transformers can be directly used with EEG data, thereby driving improvements to the classic Transformer theory.[198] Spatially, EEG processing is transferred to vision Transformers to guide the compression and evaluation of feature maps to establish dependencies.[196,205] With the discovery of more intrinsically relevant features, it is hoped that Transformers reveal more EEG features and become the

TABLE 6 Details and quantitative assessments of Transformers in EEG processing.

Author	Application	Year	Loss	Optimizer	Accuracy (%)	95HD (mm)
Wan et al.[193]	Target frequency identification	2023	CE	Adam	70.15	-
	Emotion recognition		L1		91.58	
	Depression discrimination				72.19	
Xie et al.[194]	Motor imagery	2022	-	Adam	82.95	-
					87.26	
Song et al.[195]	Motor imagery	2021	CE	Adam	91.30	82.37
					84.26	84.26
Lee et al.[196]	Overt speech	2022	Hinge[197]	-	49.50	-
	Imagined speech				35.07	
Liu et al.[198]	Motion recognition	2022	CE	Adam	93.10	92.61
					96.28	
					83.27	
Du et al.[199]	Person identification	2022	CE	AdamW	97.29	-
					97.45	
Wang et al.[200]	Emotion recognition	2022	CE	Adam	65.75	64.29
					66.51	66.27
Kostas et al.[201]	Motor imagery	2021	Cosine similarity CE[201]	Adam	86.70	-
	Error related negativity				42.60	
	Douchin seller					
	Sleep staging					
Tao et al.[202]	Human brain-visual	2021	CE	Adam	61.11	-
	Motor imagery				55.40	
Siddhad et al.[203]	Age classification	2022	CE	Adam	94.53	93.55
	Gender classification				87.79	87.99
Ahn et al.[204]	Motor imagery	2022	CE	AdamW	62.00	-
	Visual imagery				70.00	
	Speech imagery				72.20	
Ma et al.[205]	Motor imagery	2022	CE	Adam	83.90	78.20

Abbreviations: 95HD, 95% Hausdorff Distance; Adam, Adaptive moment estimation; AdamW, Adam Weight decay; CE, Cross Entropy; EEG, electroencephalogram.

most significant paradigm for man-machine interaction and brain-computer interfaces.

3.8 | Multi-task collaboration

With the continuous expansion of the parameter scale, a single deep learning model has the potential to eventually undertake the coordination of diverse tasks. Taking advantage of the adequate learning space gained from using stacked multi-head attention, some studies have designed Transformers for collaborative multitasking. For example, considering the different semantic associations of MRI sampling in different directions, Jun et al.[212] introduced a medical Transformer using multi-view embedding to enrich input features. They sampled MRI sequences from axial, sagittal, and coronal directions and completed vector extraction in a pre-trained convolutional encoder. Subsequently, the Transformer was adopted to achieve self-force enhancement in different directions, applicable in the prediction of brain disease, brain age, and brain tumor segmentation. Li et al.[213] proposed integration of CNN and Transformer architectures (Trans-ResNet). They adopted a reliable gradient transfer to achieve the efficient feature learning of attention modules to estimate age and infer AD predictions. With the development of the generalization model and the exploration of potential space, there will be greater demand for collaborative tasks, and multi-task Transformers will find use in diverse applications.

4 | OPEN SOURCE

With the improvement of ethical principles for the use of open-source resources and the development of advanced deep learning frameworks, open resources facilitate the application of Transformers in brain science. Here, we compile some of the representative publicly available datasets and toolkits for brain science research.

4.1 | Datasets

The Brain Tumor Segmentation (BraTS) challenge is one of the most popular competitions run by Medical Image Computing and Computer Assisted Intervention (MICCAI). The dataset presents four modes as T1, T1-CE, T2, and FLAIR volumes and specifies the necrotic (NCR) parts, the peritumoral edematous/invaded tissue (ED), and Enhancing Tumor (ET). The aim is to focus on the shape, appearance and histology of brain tumors, leading to the exploration of the advanced segmentation method for brain tumors. The latest BraTS 2023 is available at: https://www.synapse.org/#!Synapse:syn51156910/wiki/622351.

The Healthy MR database is from Centre of Advanced Studies and Innovation Lab at the University of North Carolina at Chapel Hill. It contains the brain images of 100 healthy subjects, ranging in age from 18 to 60+ years and is equally divided by gender. The data exclude psychiatric disease, head trauma, diabetes, hypertension and other symptoms/history, presenting four modes, namely T1, T2, MRA, and DT1. The aim is to understand the range of shapes of healthy anatomy so that statistical analyses can be defined to assess disease. It is now available at: https://data.kitware.com/#collection/591086ee8d777f16d01e0724/folder/58a372e38d777f0721a64dc6.

The Information eXtraction from Images (IXI) brain development dataset is jointly collected by three different hospitals in London and contains nearly 600MR images of healthy subjects. The collection protocols for each subject include T1, T2, Proton Density (PD), MRA, and DWI. The dataset is available at http://brain-development.org/ixi-dataset/.

The Go-nogo categorization and detection task (OpenNeuro) dataset is collected by Stanford University. It contains 32-channel EEG data obtained from 14 subjects, including seven males and seven females. All participants naturally performed 2500 image classification and detection tasks and were sampled during at least 1000 ms. The dataset is available at https://openneuro.org/datasets/ds002680/versions/1.2.0.

The Cerebral Aneurysm Detection Dataset (CADA) comprises about 200 images of cerebral aneurysms without vasospasm. All patients receive a manual injection of contrast agent into the anterior aneurysm or posterior aneurysms artery and are scanned with digital subtraction. The area of interest is selected by a neurosurgeon and the annotator provides the aneurysm segmentation. This dataset is intended to enable the early detection of aneurysms and determine prevention strategies. Details of the dataset are available at https://cada.grand-challenge.org/Dataset/.

4.2 | Toolboxes

AllenNLP is an open-source library for the deep learning of NLP established by the Allen Institute for Artificial Intelligence. It was developed based on PyTorch implementation and provides common builds and models, including BERT and GPT, which can easily be extended to custom Transformers. AllenNLP is available at https://github.com/allenai/allennlp.

TorchIO is a compact modular library based on PyTorch implementation that can be adapted to read/write port, resampling, preprocessing, and enhancement of 3D medical images for deep learning workflows, including the latest Transformers. The library is suitable for the repeatability and traceability of experiments, providing convenience for researchers in the preprocessing pipeline of complex medical image processing. It is now available at https://github.com/fepegar/torchio.

Medical Open Network for Artificial Intelligence (MONAI) was developed jointly by NVIDIA and King's College London. It exploits an open-source medical image

framework accelerated by NVIDIA technology and implements popular models including various advanced Transformers. Due to its modular design, the MONAI enables researchers to flexibly define deep learning development and utilizes code for preprocessing and transformation in the AI pipeline, thus providing consistent standards for developers and cloud providers. MONAI is now available at https://github.com/Project-MONAI/MONAI.

5 | DISCUSSION

5.1 | Quantitative assessment

In this subsection, we summarize the quantitative assessments of several of the prominent Transformers in brain sciences. These results are widely discussed in visual missions or Transformer design. We first extend the quantitative results of brain disease diagnosis in Table 1, which shows that brain tumor classification has become the most representative exploration owing to its diversified characterization in multi-modal brain images. Transformers may provide a vital reference for integrated auxiliary diagnosis research. We have also added assessments of brain age prediction in Table 2, which shows that such Transformers are always trained using rich samples to better fit the regression curve.

We expand on the quantitative assessments of Transformers in brain tumor segmentation in Table 3. Most of these models were assessed on BraTS datasets from MICCAI, being state-of-the-art or nearly so. Owing to class distribution imbalance and structural complexity, cerebrovascular segmentation is more challenging (see Table 4), indicating more opportunities for innovation in Transformers. In contrast, brain tissue segmentation demonstrates capability even in small sample datasets owing to its more significant semantic features (see Table 5). Finally, Table 6 presenting EEG processing results shows that motor imagery is the most typical application and outperforms other applications. It is expected that more novel Transformers will be used for the interpretation of EEG signals from complex behaviors, such as emotion recognition, overt speech, and imagined speech.

In summary, owing to the superior properties of multi-head attention, extensibility, and global dependence, we have witnessed a significant improvement in deep learning with the development of Transformers. Although some brain sciences with Transformers are a preliminary attempt, we believe that they are expected to reach the state-of-the-art with the enrichment of datasets and the proposal of more advanced theories.

5.2 | Model complexity

As discussed in Section 5.1, the excellent spatial expansion of Transformers drives their capability to model complex tasks. While this attribute is a notable feature of Transformers, it also raises wide concerns about model complexity in Transformers. In Table 7, we have collected and recorded the model complexity of diverse Transformers that have been publicly reported in brain sciences. It shows that some Transformers have high inference and computational costs, for example, the mmFormer[138] reported the parameters and FLOating Point operations (FLOPs) with 106.00M and 748G, which required a large computational memory (25G stated in Ref. [138]) to complete the model computation. Hatamizadeh et al.[19,125] proposed the UNETR and Swin UNETR successively. Although the Swin Transformer has contributed in parameter weight, it still faces computational complexity. In addition, Chen et al.[147] presented a complexity comparison between Transformers and representative CNNs, which demonstrates that Transformers have more significant model complexity than most CNNs.

Recently, large Transformers have been rapidly developed, with Chat Generative Pre-trained Transformer (ChatGPT) as an example that contains over 175 billion parameters to achieve stronger logical thinking and reasoning skills.[214,215] The latest version of ChatGPT has successfully passed the United States Medical Licensing Examination,[216] showing significant potential for clinical application. To integrate medical image processing capabilities in ChatGPT, Chat Computer-Aided Diagnosis (ChatCAD) exploited the large language model for computer-aided diagnosis.[217] It generated vital prompt texts (i.e., diagnostics information, lesion segmentation and pathological reports) that can be transferred to the large

TABLE 7 Model complexity for different Transformers.

Method	Param. (M)	FLOPs (G)
TransBTS[38]	32.99	333.09
TransforMesh[83]	4.10	-
NestedFormer[137]	10.48	71.77
Swin UNETR[125]	61.98	394.84
BiTr-Unet[131]	-	-
TransConver[128]	9.00	66.70
mmFormer[138]	106.00	748.00
TransBTSV2[129]	15.30	240.66
UNETR[19]	92.58	41.19
VT-UNet[134]	<20.00	101.00
BTSwin-Unet[132]	35.60	69.10
DenseTrans[135]	21.30	212.00
3D PSwinBTS[126]	20.40	68.60
TRSF-Net[147]	-	274.09
SF-SegFormer[153]	1.00	-
UNesT[155]	87.30	261.70
TW-Net[156]	30.09	5.81

Abbreviation: FLOPs, FLOating Point operations.

language model to present professional medical suggestions. ChatCAD employed 227,835 radiographic studies and images from 65,240 patients in training for just one application scenario. By integrating the large image and prompt encoders, the universal Segment Anything Model (SAM) achieved promising zero-sample generalization.[218] SAM was trained using the largest segment dataset (i.e., the Anything 1-Billion mask Dataset), with a total of about 11 million images and 1 billion masks. Rich features enable SAM to have robust semantic awareness and generalization. To apply it to medical images, the research field has developed SAM compression and distillation techniques, such as MedSAM,[219] which leveraged a dataset with 200,000 image masks and 11 modalities to successfully distillate SAM into universal medical image segmentation. The improvement of these large Transformers provides more valuable recommendations and decisions for the brain sciences. These advanced Transformers also encourage more thinking and attention to model complexity.

Although the development of Swin Transformers[128,134] and windowed attention[153] has moderated computational complexity, the need for more efficient attention mechanisms remains. Fortunately, many mature operators have been developed in CNNs,[220,221] which are beneficial to provide theoretical guidance for more efficient computational mechanisms. Some advanced cloud computing has been successfully applied to medical image segmentation, which helps alleviate the imbalance of computing resources and contribute large model computation. This will actively promote the development of Transformers in the calculation mechanism.

5.3 | Optimization

We performed statistical analysis of the optimization of Transformers in specific scenarios for the convenience of researchers to review and further reproduce. It shows that the most classical optimization scheme is the combination of CE loss and Adam, which has been widely proven to be conducive to fast convergence.[222,223] When samples are not evenly distributed (i.e., cerebrovascular and brain tumor), the weight CE loss is a classical scheme to balance target constraints and has developed Cosine similarity[201] and Hessian soft weights.[148] In a segmentation scenario, the L2 and Dice regularization are often combined with CE loss to supplement empirical priors and reduce overfitting due to excessive predicted voxels. Moreover, some researches have proposed proprietary regularization, such as the FOG,[157] topological,[158] and Structural SIMilarity (SSIM) loss,[143] to be adapted to specific tasks. The MSE and MAE losses are often used for regression (i.e., brain age estimation) to match the distribution of the real sample.

Despite their simplicity and convenience, the Stochastic Gradient Descent (SGD) and Momentum may be rarely used in large-scale model training owing to their tendency to fall into local minimums.[224] Compared to the Adam optimizer, the AdamW[72] incorporates the gradient of the regularization term into the backpropagation for more efficient computation. Its advantages are predicted to become an important propellant for the development of large-scale Transformers, and these advanced optimizations will be an important internal mechanism for Transformers to achieve their target effect.

6 | CHALLENGES AND OPPORTUNITIES

Although Transformers have exhibited remarkable performance in diverse applications, some challenges remain, such as high computational complexity, heavy parameters and inflexible data conversion between CNNs and Transformers, inspiring some expected development opportunities. We have reviewed the existing challenges and attempted to summarize some opportunities for in-depth discussion and enlightenment.

6.1 | Graph transformers

To facilitate large-scale and deep data storage and normalization, structured data, such as image, volume, and sequence, are the most common manifestation in medicine. Owing to sophisticated computational theory and simple processing, structured data have become the typical pattern in brain sciences exploration. However, structured data lack flexible and adaptable modeling for references, which can create limitations in describing the complex structure and function of the brain.

As a representative of unstructured data, the graph is composed of vertex and edge sets, which describes the dependencies within the object. According to the latest work on Transformers, graph-based modeling has also gradually been developed; for example, the Graph Transformer[98] exploited the brain template of prior knowledge to calculate RoIs and construct them as graphs. TRSF-Net[147] and GraformerDIR[170] model diverse spatial distributions from feature maps as various graphs. These Transformers completed geometry learning and captured high-quality dependencies within these models. We look forward to the Graph Transformers of the near future, with more flexible modeling and suitability to long-distance dependence learning, for comprehensive and complicated brain science tasks, such as those involving neurons, functional regions, and RoIs.

6.2 | Vision-based transformers

Since most works have focused on NLP-based Transformers, integration and modularization for the attention module are the most common applications. Especially in visual tasks, research established the embedding layer for flattening feature maps as tokens. For example, the typical studies based on ViT extensively convert the image patches into

vectors using fully connected layers.[225,226] Owing to their rich theoretical basis, NLP-based Transformers are straightforward and popular but may have some limitations. For example, compacting feature space with rich semantics into tokens is destructive to semantic integrity. In addition, the autoregressive property of NLP-Transformers is rarely reflected in visual tasks. We have noticed that the latest works have gradually captured tokens from RoIs, especially some tokens learned from channels[227] and structured space.[228] Visual-based Transformers need to deduce and reproduce complicated mechanisms (such as self-attention and multi-head attention), generating the 2D or even 3D vectors that are dedicated to visual-based Transformers. It is likely that vision-based Transformers will attract more attention, although these Transformers may lack uniformity in vision and NLP, which may induce a new discussion on the tradeoff between uniformity and adaptability in vision.

6.3 | Prompt-based transformers

Prompt-based learning is a novel paradigm in deep learning.[229] In Transformers, models have gradually expanded from the pre-train and fine-tuning paradigm to prompt-based learning for different downstream tasks.[230] A study found that prompt learning can significantly improve the representation learning ability of ViTs because it is proven to integrate domain knowledge effectively.[231] Another work demonstrated that prompt integration into learnable parameters can achieve excellent performance in medical image segmentation.[232]

Transformers optimize the embedding and task headers in the specific task via prompt-based learning, which guides the distribution of training data close to the task description. Thus, Transformers pay more attention to guidance, pursuing training efficiency and accuracy. Owing to the beneficial attributes of awareness consistency and few- or zero-shot learning, prompt-based Transformers can embed more specialized domain semantics and even guide users to add prior knowledge and are expected to attract more attention for the multi-modal processing of knowledge and data.

6.4 | Hyper-heavy transformers

Model extensibility research proved that Transformers can gradually gain capabilities with the expansion of datasets and model parameters,[233,234] primarily due to the presentation capabilities of the Transformer and its variant architectures, especially the multi-head attention module, which maps the original information to subspaces to capture diverse potential features. Due to the abandonment of the nature of translation invariability and weight sharing like CNNs, the expansion of Transformers is prone to generate hyper-heavy parameters.

We have noticed that recent hyper-heavy Transformers have performed optimally, which is a testament to their outstanding potential.[214,215,218] While they inevitably lead to significantly increased computing requirements and processing times, such Transformers are conducive to explain more complex application scenarios. In addition, the hyper-heavy Transformer is also a vital pre-training model that can assist more detailed downstream tasks, such as semantic segmentation and object detection and classification. While we expect the development of more advanced hyper-heavy Transformers for realistic and complicated scenarios, this stage is still out of reach for most researchers. Therefore, model compression and fine-tuning techniques are expected to be the next main direction. The latest research has successfully implemented hyper-heavy Transformers for mission-specific applications by freezing part of the parameters and fine-tuning[217,219]; this may attract more exploration in the near future.

6.5 | Lightweight transformers

Although hyper-heavy Transformers have performed optimally, they also significantly increase training and deployment costs. Extensible research evidence has proved that hyper-heavy Transformers are mainly adapted to large-dataset learning.[233,234] Therefore, recently proposed Transformers[128,153] comprise lightweight models to better adapt to user-customized datasets and terminals, especially in the deployment of mobile terminal, blockchain, and the Internet of Things (IoTs).

As discussed in Section 5.2, Transformers always load heavy parameters because of stacked attention modules and multi-head attention mechanics. Recent works have focused on some attempts to develop lightweight Transformers; for example, Chen[147] and Xing[137] et al. shortened the attention modules in the case of sufficient semantics. In addition, Swin Transformers contribute to reducing parameters owing to their shift window-based attention. However, these methods either sacrifice accuracy or rely on shift windows, which may be affected when establishing dependence over long distances. Therefore, there is an urgent requirement for more efficient attention mechanisms to adapt to feature subspace in multi-head attention, such as a priori-guided learning space,[221] the compact computing module,[220] and the sparse matrix or attention.[235] Such methods will reduce redundant attention parameters, separate coupled features, and model sparse matrices, which will promote the development of lightweight Transformers.

6.6 | Open medical datasets

Due to the impressive features of the latest large-scale Transformers,[217,218] recent efforts have explored Transformer-based medical large-scale models to enable more accurate clinical analysis and decision-making. For example, the latest AutoSAM attempts to train MRI with few labels via few-shot learning, with outstanding results on public medical image segmentation datasets.[236] Although advanced compression and fine-tuning techniques have been applied to a

variety of mission-specific scenarios, the research community is still looking forward to the birth of medical large models, which will contain richer medical domain-specific semantics to provide more accurate predictions.

Unlike general semantic image tasks, medical data require perfect ethics and privacy protection, a major challenge in the development of large-scale open medical datasets. The statistical application scenarios in this paper show that there is still a lack of unified data normalization in brain sciences, causing a lack of robust generalization. Fortunately, there is also growing consensus about specific medical tasks; for example, MICCAI updates the BraTS dataset every year to provide new challenges and opportunities.[237] References[238] also attempt to build simulation data using generative adversarial models for privacy protection. As medical data statements are refined, it is expected that more datasets will be made publicly available to researchers. Highly accurate annotations and more robust data augmentation theories will be urgently needed.

7 | CONCLUSIONS

Transformers have presented superior performance owing to their strong scaling and long-distance dependence. This review provides a comprehensive overview of the use of Transformers in brain sciences. We focused on discussing these advanced Transformers regarding data modality, application, loss function, optimization, and quantitative results. These details may support researchers in reproduction and further improvement. We hope that this review provides a comprehensive exploration of Transformers in brain sciences that will inspire further thinking and innovation in brain sciences and related research fields.

AUTHOR CONTRIBUTIONS

Cheng Chen: Conceptualization; methodology; project administration; writing – original draft. **Huilin Wang**: Methodology; visualization; writing – original draft. **Yunqing Chen**: Project administration; visualization; writing – original draft. **Zihan Yin**: Methodology; resources; writing – original draft. **Xinye Yang**: Formal analysis; writing – original draft. **Huansheng Ning**: Supervision; writing – review & editing. **Qian Zhang**: Resources; writing – review & editing. **Weiguang Li**: Resources; writing – review & editing. **Ruoxiu Xiao**: Funding acquisition; writing – review & editing. **Jizong Zhao**: Resources; supervision; writing – review & editing.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Natural Science Foundation of China (62176268), China Postdoctoral Science Foundation (2023M730226), Nonprofit Central Research Institute Fund of Chinese Academy of Medical Sciences (2020-JKCS-008), Major Science and Technology Project of Zhejiang Province Health Commission (WKJ-ZJ-2112), and Fundamental Research Funds for the Central Universities (FRF-TP-22-047A1).

CONFLICT OF INTEREST STATEMENT

Jizong Zhao is a member of the Editorial Board for Brain-X. The manuscript was handled by another Editor and has undergone a rigorous peer-review process. Jizong Zhao was not involved in the journal's review of/or decision related to this manuscript. The other authors declare that they have no conflicts of interest.

ORCID

Cheng Chen https://orcid.org/0000-0002-4203-2145

REFERENCES

1. Konstantinides N, Holguera I, Rossi AM, et al. A complete temporal transcription factor series in the fly visual system. *Nature*. 2022;604(7905):316-322. https://doi.org/10.1038/s41586-022-04564-w
2. Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. Backpropagation and the brain. *Nat Rev Neurosci*. 2020;21(6):335-346. https://doi.org/10.1038/s41583-020-0277-3
3. Miller JA, Ding S-L, Sunkin SM, et al. Transcriptional landscape of the prenatal human brain. *Nature*. 2014;508(7495):199-206. https://doi.org/10.1038/nature13185
4. Oh SW, Harris JA, Ng L, et al. A mesoscale connectome of the mouse brain. *Nature*. 2014;508(7495):207-214. https://doi.org/10.1038/nature13186
5. Hsu W-Y. Brain–computer interface: the next frontier of telemedicine in human–computer interaction. *Telematics Inf*. 2015;32(1):180-192. https://doi.org/10.1016/j.tele.2014.07.001
6. Sung SH, Jeong Y, Oh JW, Shin H-J, Lee JH, Lee KJ. Bio-plausible memristive neural components towards hardware implementation of brain-like intelligence. *Mater Today*. 2023;62:251-270. https://doi.org/10.1016/j.mattod.2022.11.022
7. Rong G, Mendez A, Assi EB, Zhao B, Sawan M. Artificial intelligence in healthcare: review and prediction case studies. *Engineering*. 2020;6(3):291-301. https://doi.org/10.1016/j.eng.2019.08.015
8. Rezeika A, Benda M, Stawicki P, Gembler F, Saboor A, Volosyak I. Brain–computer interface spellers: a review. *Brain Sci*. 2018;8(4):57. https://doi.org/10.3390/brainsci8040057
9. D'Esposito M, Postle BR. The cognitive neuroscience of working memory. *Annu Rev Psychol*. 2015;66(1):115-142. https://doi.org/10.1146/annurev-psych-010814-015031
10. Arpaia P, De Benedetto E, Duraccio L. Design, implementation, and metrological characterization of a wearable, integrated AR-BCI hands-free system for health 4.0 monitoring. *Measurement*. 2021;177:109280. https://doi.org/10.1016/j.measurement.2021.109280
11. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*; 2017.
12. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. *AI Open*. 2022;3:111-132. https://doi.org/10.1016/j.aiopen.2022.10.001
13. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. *ACM Comput Surv*. 2022;54(10s):1-41. https://doi.org/10.1145/3505244
14. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018: arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805
15. Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training. 2018.
16. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. *OpenAI Blog*. 2019;1(8):9.
17. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: *Computer Vision – ECCV 2020*. Springer; 2020:213-229.
18. Misra I, Girdhar R, Joulin A. An end-to-end transformer model for 3D object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021:2906-2917.

19. Hatamizadeh A, Tang Y, Nath V, et al. Unetr: transformers for 3D medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; 2022:574-584.
20. Xiao H, Li L, Liu Q, Zhu X, Zhang Q. Transformers in medical image segmentation: a review. *Biomed Signal Process Control*. 2023;84:104791. https://doi.org/10.1016/j.bspc.2023.104791
21. Chen C, Zhou K, Wang Z, Zhang Q, Xiao R. All answers are in the images: a review of deep learning for cerebrovascular segmentation. *Comput Med Imag Graph*. 2023;107:102229. https://doi.org/10.1016/j.compmedimag.2023.102229
22. Das S, Nayak GK, Saba L, Kalra M, Suri JS, Saxena S. An artificial intelligence framework and its bias for brain tumor segmentation: a narrative review. *Comput Biol Med*. 2022;143:105273. https://doi.org/10.1016/j.compbiomed.2022.105273
23. Puffay C, Accou B, Bollens L, et al. Relating EEG to continuous speech using deep neural networks: a review. 2023:arXiv:2302.01736. https://doi.org/10.48550/arXiv.2302.01736
24. Tanveer M, Ganaie MA, Beheshti I, et al. Deep learning for brain age estimation: a systematic review. *Inf Fusion*. 2023;96:130-143. https://doi.org/10.1016/j.inffus.2023.03.007
25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016:770-778.
26. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2015:3431-3440.
27. Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans Neural Netw Learn Syst*. 2022;33(12):6999-7019. https://doi.org/10.1109/TNNLS.2021.3084827
28. Holt CE, Martin KC, Schuman EM. Local translation in neurons: visualization and function. *Nat Struct Mol Biol*. 2019;26(7):557-566. https://doi.org/10.1038/s41594-019-0263-5
29. Cohen MX. Where does EEG come from and what does it mean? *Trends Neurosci*. 2017;40(4):208-218. https://doi.org/10.1016/j.tins.2017.02.004
30. Chen P, Hong W. Neural circuit mechanisms of social behavior. *Neuron*. 2018;98(1):16-30. https://doi.org/10.1016/j.neuron.2018.02.026
31. Rao RM, Liu J, Verkuil R, et al. MSA transformer. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR; 2021:8844-8856.
32. Liu Z, Ning J, Cao Y, et al. Video swin transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022:3202-3211.
33. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379(6637):1123-1130. https://doi.org/10.1126/science.ade2574
34. Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J Mach Learn Res*. 2022;23(1):5232-5270.
35. Wang D, Zhang Q, Xu Y, et al. Advancing plain vision transformer towards remote sensing foundation model. *IEEE Trans Geosci Rem Sens*. 2023;61:1-15. https://doi.org/10.1109/TGRS.2022.3222818
36. Shamshad F, Khan S, Zamir SW, et al. Transformers in medical imaging: a survey. *Med Image Anal*. 2023;88:102802. https://doi.org/10.1016/j.media.2023.102802
37. Xu Y, Wei H, Lin M, et al. Transformers in computational visual media: a survey. *Comput Vis Media*. 2022;8(1):33-62. https://doi.org/10.1007/s41095-021-0247-3
38. Wang W, Chen C, Ding M, Yu H, Zha S, Li J. Transbts: multimodal brain tumor segmentation using transformer. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer; 2021:109-119.
39. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015:arXiv:1511.06434. https://doi.org/10.48550/arXiv.1511.06434
40. Selva Birunda S, Kanniga Devi R. A review on word embedding techniques for text classification. In: *Innovative Data Communication Technologies and Application*; 2021:267-281.
41. Okada S, Ohzeki M, Taguchi S. Efficient partition of integer optimization problems with one-hot encoding. *Sci Rep*. 2019;9(1):13036. https://doi.org/10.1038/s41598-019-49539-6
42. Rodríguez P, Bautista MA, Gonzalez J, Escalera S. Beyond one-hot encoding: lower dimensional target embedding. *Image Vis Comput*. 2018;75:21-31. https://doi.org/10.1016/j.imavis.2018.04.004
43. Lai S, Liu K, He S, Zhao J. How to generate a good word embedding. *IEEE Intell Syst*. 2016;31(6):5-14. https://doi.org/10.1109/MIS.2016.45
44. Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021:12179-12188.
45. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. 2020:arXiv:2010.11929. https://doi.org/10.48550/arXiv.2103.13915
46. Parvaiz A, Khalid MA, Zafar R, Ameer H, Ali M, Fraz MM. Vision transformers in medical computer vision—a contemplative retrospection. *Eng Appl Artif Intell*. 2023;122:106126. https://doi.org/10.1016/j.engappai.2023.106126
47. Xu J, Moyer D, Grant PE, Golland P, Iglesias JE, Adalsteinsson E. SVoRT: iterative transformer for slice-to-volume registration in fetal brain MRI. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer; 2022:3-13.
48. Chan ER, Lin CZ, Chan MA, et al. Efficient geometry-aware 3D generative adversarial networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022:16123-16133.
49. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021:10012-10022.
50. Sutskever I, Martens J, Hinton GE. Generating text with recurrent neural networks. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*; 2011:1017-1024.
51. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Process*. 1997;45(11):2673-2681. https://doi.org/10.1109/78.650093
52. Ke G, He D, Liu T-Y. Rethinking positional encoding in language pre-training. 2020:arXiv:2006.15595. https://doi.org/10.48550/arXiv.2006.15595
53. Pham N-Q, Ha T-L, Nguyen T-N, et al. Relative positional encoding for speech recognition and direct translation. 2020:arXiv:2005.09940. https://doi.org/10.48550/arXiv.2005.09940
54. Chen P-C, Tsai H, Bhojanapalli S, Chung HW, Chang Y-W, Ferng C-S. A simple and effective positional encoding for transformers. 2021:arXiv:2104.08698. https://doi.org/10.48550/arXiv.2104.08698
55. Chu X, Tian Z, Zhang B, et al. Conditional positional encodings for vision transformers. 2021:arXiv:2102.10882. https://doi.org/10.48550/arXiv.2102.10882
56. Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014:arXiv:1406.1078. https://doi.org/10.48550/arXiv.1406.1078
57. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014:arXiv:1412.3555. https://doi.org/10.48550/arXiv.1412.3555
58. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
59. Chen C, Zhan L, Pan X, et al. Automatic recognition of auditory brainstem response characteristic waveform based on bidirectional long short-term memory. *Front Med*. 2021;7:613708. https://doi.org/10.3389/fmed.2020.613708

60. Joulin A, Cissé M, Grangier D, Jégou H. Efficient softmax approximation for GPUs. In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR; 2017:1302-1310.
61. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings*; 2011:315-323.
62. Tetko IV, Karpov P, Van Deursen R, Godin G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat Commun*. 2020;11(1):5575. https://doi.org/10.1038/s41467-020-19266-y
63. Mahmud T, Marculescu D. AVE-CLIP: AudioCLIP-based multi-window temporal transformer for audio visual event localization. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; 2023:5158-5167.
64. Alshemali B, Kalita J. Improving the reliability of deep neural networks in NLP: a review. *Knowl Base Syst*. 2020;191:105210. https://doi.org/10.1016/j.knosys.2019.105210
65. Liu J, Chen F, Wang X, et al. A comparative analysis framework of 3T and 7T TOF-MRA based on automated cerebrovascular segmentation. *Comput Med Imag Graph*. 2021;89:101830. https://doi.org/10.1016/j.compmedimag.2020.101830
66. Schültke E, Fiedler S, Nemoz C, et al. Synchrotron-based intravenous K-edge digital subtraction angiography in a pig model: a feasibility study. *Eur J Radiol*. 2010;73(3):677-681. https://doi.org/10.1016/j.ejrad.2009.01.019
67. Mehanna R, Jankovic J. Movement disorders in cerebrovascular disease. *Lancet Neurol*. 2013;12(6):597-608. https://doi.org/10.1016/S1474-4422(13)70057-7
68. Chen C, Zhou K, Guo X, Wang Z, Xiao R, Wang G. Cerebrovascular segmentation in phase-contrast magnetic resonance angiography by multi-feature fusion and vessel completion. *Comput Med Imag Graph*. 2022;98:102070. https://doi.org/10.1016/j.compmedimag.2022.102070
69. Lerch JP, Van Der Kouwe AJ, Raznahan A, et al. Studying neuroanatomy using MRI. *Nat Neurosci*. 2017;20(3):314-326. https://doi.org/10.1038/nn.4501
70. Wadhwa A, Bhardwaj A, Singh Verma V. A review on brain tumor segmentation of MRI images. *Magn Reson Imaging*. 2019;61:247-259. https://doi.org/10.1016/j.mri.2019.05.043
71. Odusami M, Maskeliūnas R, Damaševičius R. Pixel-level fusion approach with vision transformer for early detection of Alzheimer's disease. *Electronics*. 2023;12(5):1218. https://doi.org/10.3390/electronics12051218
72. Huang Y, Zhou J, Li X, et al. MENet: map-enhanced 3D object detection in bird's-eye view for LiDAR point clouds. *Int J Appl Earth Obs Geoinf*. 2023;120:103337. https://doi.org/10.1016/j.jag.2023.103337
73. Shin H, Jeon S, Seol Y, Kim S, Kang D. Vision transformer approach for classification of Alzheimer's disease using 18F-florbetaben brain images. *Appl Sci*. 2023;13(6):3453. https://doi.org/10.3390/app13063453
74. Tummala S, Kadry S, Bukhari SAC, Rauf HT. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Curr Oncol*. 2022;29(10):7498-7511. https://doi.org/10.3390/curroncol29100590
75. McMahan B, Streeter M. Delay-tolerant algorithms for asynchronous distributed online learning. In: *Advances in Neural Information Processing Systems*; 2014.
76. Zeiler MD. ADADELTA: an adaptive learning rate method. 2012: arXiv:1212.5701. https://doi.org/10.48550/arXiv.1212.5701
77. Li Z, Cong Y, Chen X, et al. Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. *iScience*. 2023;26(1):105872. https://doi.org/10.1016/j.isci.2022.105872
78. Qiu Y, Yu S, Zhou Y, et al. Multi-channel sparse graph transformer network for early Alzheimer's disease identification. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*; 2021:1794-1797.
79. Aloraini M, Khan A, Aladhadh S, Habib S, Alsharekh MF, Islam M. Combining the transformer and convolution for effective brain tumor classification using MRI images. *Appl Sci*. 2023;13(6):3680. https://doi.org/10.3390/app13063680
80. Zhou X, Tang C, Huang P, Tian S, Mercaldo F, Santone A. ASI-DBNet: an adaptive sparse interactive resnet-vision transformer dual-branch network for the grading of brain cancer histopathological images. *Interdiscip Sci*. 2023;15(1):15-31. https://doi.org/10.1007/s12539-022-00532-0
81. Anaya-Isaza A, Mera-Jiménez L, Verdugo-Alejo L, Sarasti L. Optimizing MRI-based brain tumor classification and detection using AI: a comparative analysis of neural networks, transfer learning, data augmentation, and the cross-transformer network. *Eur J Radiol Open*. 2023;10:100484. https://doi.org/10.1016/j.ejro.2023.100484
82. Ferdous GJ, Sathi KA, Hossain MA, Hoque MM, Dewan MAA. LCDEiT: a linear complexity data-efficient image transformer for MRI brain tumor classification. *IEEE Access*. 2023;11:20337-20350. https://doi.org/10.1109/ACCESS.2023.3244228
83. Sarasua I, Pölsterl S, Wachinger C. TransforMesh: a transformer network for longitudinal modeling of anatomical meshes. In: *Machine Learning in Medical Imaging*. Springer International Publishing; 2021:209-218.
84. Samak ZA, Clatworthy P, Mirmehdi M. TranSOP: transformer-based multimodal classification for stroke treatment outcome prediction. 2023:arXiv:2301.10829. https://doi.org/10.48550/arXiv.2301.10829
85. Zhou J, Muirhead W, Williams SC, Stoyanov D, Marcus HJ, Mazomenos EB. Shifted-windows transformers for the detection of cerebral aneurysms in microsurgery. *Int J Comput Assist Radiol Surg*. 2023;18(6):1-9. https://doi.org/10.1007/s11548-023-02871-9
86. Ebrahimighahnavieh MA, Luo S, Chiong R. Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. *Comput Methods Progr Biomed*. 2020;187:105242. https://doi.org/10.1016/j.cmpb.2019.105242
87. Shoeibi A, Khodatars M, Jafari M, et al. Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: a review. *Inf Fusion*. 2023;93:85-117. https://doi.org/10.1016/j.inffus.2022.12.010
88. Lv Z, Mi F, Wu Z, et al. A parallel cerebrovascular segmentation algorithm based on focused multi-Gaussians model and heterogeneous Markov random field. *IEEE Trans Nanobiosci*. 2020;19(3):538-546. https://doi.org/10.1109/TNB.2020.2996604
89. Chen C, Xiao R, Zhang T, et al. Pathological lung segmentation in chest CT images based on improved random walker. *Comput Methods Progr Biomed*. 2021;200:105864. https://doi.org/10.1016/j.cmpb.2020.105864
90. Frisoni GB, Boccardi M, Barkhof F, et al. Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. *Lancet Neurol*. 2017;16(8):661-676. https://doi.org/10.1016/S1474-4422(17)30159-X
91. Ossenkoppele R, Rabinovici GD, Smith R, et al. Discriminative accuracy of [18F] flortaucipir positron emission tomography for Alzheimer disease vs other neurodegenerative disorders. *JAMA*. 2018;320(11):1151-1162. https://doi.org/10.1001/jama.2018.12917
92. Guo J, Qiu W, Li X, Zhao X, Guo N, Li Q. Predicting Alzheimer's disease by hierarchical graph convolution from positron emission tomography imaging. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE; 2019:5359-5363.
93. Daerr S, Brendel M, Zach C, et al. Evaluation of early-phase [18F]-florbetaben PET acquisition in clinical routine cases. *NeuroImage Clin*. 2017;14:77-86. https://doi.org/10.1016/j.nicl.2016.10.005
94. Franke K, Gaser C. Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained? *Front Neurol*. 2019;10:789. https://doi.org/10.3389/fneur.2019.00789
95. Cearns M, Hahn T, Baune BT. Recommendations and future directions for supervised machine learning in psychiatry. *Transl Psychiatry*. 2019;9(1):271. https://doi.org/10.1038/s41398-019-0607-2
96. Mishra S, Beheshti I, Khanna P. A review of neuroimaging-driven brain age estimation for identification of brain disorders and health

conditions. *IEEE Rev Biomed Eng*. 2023;16:371-385. https://doi.org/10.1109/RBME.2021.3107372

97. He S, Grant PE, Ou Y. Global-local transformer for brain age estimation. *IEEE Trans Med Imag*. 2021;41(1):213-224. https://doi.org/10.1109/TMI.2021.3108910
98. Cai H, Gao Y, Liu M. Graph Transformer geometric learning of brain networks using multimodal MR images for brain age estimation. *IEEE Trans Med Imag*. 2023;42(2):456-466. https://doi.org/10.1109/TMI.2022.3222093
99. Hu Y, Wang H, Li B. SQET: squeeze and excitation transformer for high-accuracy brain age estimation. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2022:1554-1557.
100. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2018:7132-7141.
101. Wang W, Chen W, Qiu Q, et al. CrossFormer++: a versatile vision transformer hinging on cross-scale attention. 2023:arXiv:2303.06908. https://doi.org/10.48550/arXiv.2303.06908
102. Armanious K, Abdulatif S, Shi W, et al. Age-net: an MRI-based iterative framework for brain biological age estimation. *IEEE Trans Med Imag*. 2021;40(7):1778-1791. https://doi.org/10.1109/TMI.2021.3066857
103. Cole JH. Neuroimaging-derived brain-age: an ageing biomarker? *Aging*. 2017;9(8):1861-1862. https://doi.org/10.18632/aging.101286
104. Soomro TA, Zheng L, Afifi AJ, et al. Image segmentation for MR brain tumor detection using machine learning: a Review. *IEEE Rev Biomed Eng*. 2023;16:70-90. https://doi.org/10.1109/RBME.2022.3185292
105. Sánchez FL, Hupont I, Tabik S, Herrera F. Revisiting crowd behaviour analysis through deep learning: taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Inf Fusion*. 2020;64:318-335. https://doi.org/10.1016/j.inffus.2020.07.008
106. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*; 2015.
107. Li H, Huang J, Li G, et al. View-disentangled transformer for brain lesion detection. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2022:1-5.
108. Choromanski K, Likhosherstov V, Dohan D, et al. Rethinking attention with performers. 2020:arXiv:2009.14794. https://doi.org/10.48550/arXiv.2009.14794
109. Bergmann P, Fauser M, Sattlegger D, Steger C. Uninformed students: student-teacher anomaly detection with discriminative latent embeddings. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*; 2020:4183-4192.
110. Liu W, Li R, Zheng M, et al. Towards visually explaining variational autoencoders. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020:8642-8651.
111. Salehi M, Sadjadi N, Baselizadeh S, Rohban MH, Rabiee HR. Multiresolution knowledge distillation for anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2021:14902-14912.
112. Defard T, Setkov A, Loesch A, Audigier R. Padim: a patch distribution modeling framework for anomaly detection and localization. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Springer; 2021:475-489.
113. Rippel O, Mertens P, Merhof D. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE; 2021:6726-6733.
114. Chen L, You Z, Zhang N, Xi J, Le X. Utrad: anomaly detection and localization with u-transformer. *Neural Netw*. 2022;147:53-62. https://doi.org/10.1016/j.neunet.2021.12.008
115. Da Costa PF, Dafflon J, Mendes SL, et al. Transformer-based normative modelling for anomaly detection of early schizophrenia. 2022:arXiv:2212.04984. https://doi.org/10.48550/arXiv.2212.04984
116. Pinaya WH, Tudosiu P-D, Gray R, et al. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. *Med Image Anal*. 2022;79:102475. https://doi.org/10.1016/j.media.2022.102475
117. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. *IEEE Trans Pattern Anal Mach Intell*. 2021;44(7):3523-3542. https://doi.org/10.1109/TPAMI.2021.3059968
118. Cheng H-D, Jiang XH, Sun Y, Wang J. Color image segmentation: advances and prospects. *Pattern Recogn*. 2001;34(12):2259-2281. https://doi.org/10.1016/S0031-3203(00)00149-7
119. Despotović I, Goossens B, Philips W. MRI segmentation of the human brain: challenges, methods, and applications. *Comput Math Methods Med*. 2015;2015:450341. https://doi.org/10.1155/2015/450341
120. Gordillo N, Montseny E, Sobrevilla P. State of the art survey on MRI brain tumor segmentation. *Magn Reson Imaging*. 2013;31(8):1426-1438. https://doi.org/10.1016/j.mri.2013.05.002
121. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. 2015:arXiv:1511.07122. https://doi.org/10.48550/arXiv.1511.07122
122. Fan M, Lai S, Huang J, et al. Rethinking bisenet for real-time semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2021:9716-9725.
123. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer; 2015:234-241.
124. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Springer; 2016:424-432.
125. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: swin transformers for semantic segmentation of brain tumors in MRI images. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer; 2022:272-284.
126. Liang J, Yang C, Zeng L. 3D PSwinBTS: an efficient transformer-based Unet using 3D parallel shifted windows for brain tumor segmentation. *Digit Signal Prog*. 2022;131:103784. https://doi.org/10.1016/j.dsp.2022.103784
127. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2015:1-9.
128. Liang J, Yang C, Zeng M, Wang X. TransConver: transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images. *Quant Imag Med Surg*. 2022;12(4):2397-2415. https://doi.org/10.21037/qims-21-919
129. Li J, Wang W, Chen C, et al. TransBTSV2: towards better and more efficient volumetric segmentation of medical images. 2022:arXiv:2201.12785. https://doi.org/10.48550/arXiv.2201.12785
130. Lyu Q, Namjoshi SV, McTyre E, et al. A transformer-based deep-learning approach for classifying brain metastases into primary organ sites using clinical whole-brain MRI images. *Patterns*. 2022;3(11):100613. https://doi.org/10.1016/j.patter.2022.100613
131. Jia Q, Shu H. Bitr-unet: a CNN-transformer combined network for MRI brain tumor segmentation. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer; 2022:3-14.
132. Liang J, Yang C, Zhong J, Ye X. BTSwin-Unet: 3D U-shaped symmetrical swin transformer-based network for brain tumor segmentation with self-supervised pre-training. *Neural Process Lett*. 2022;55(4):1-19. https://doi.org/10.1007/s11063-022-10919-1
133. Jiang Y, Zhang Y, Lin X, Dong J, Cheng T, Liang J. SwinBTS: a method for 3D multimodal brain tumor segmentation using swin transformer. *Brain Sci*. 2022;12(6):797. https://doi.org/10.3390/brainsci12060797
134. Peiris H, Hayat M, Chen Z, Egan G, Harandi M. A robust volumetric transformer for accurate 3D tumor segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer; 2022:162-172.

135. ZongRen L, Silamu W, Yuzhen W, Zhe W. DenseTrans: multimodal brain tumor segmentation using swin transformer. *IEEE Access*. 2023;11:42895-42908. https://doi.org/10.1109/ACCESS.2023.3272055
136. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: a nested u-net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer; 2018:3-11.
137. Xing Z, Yu L, Wan L, Han T, Zhu L. Nestedformer: nested modality-aware transformer for brain tumor segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer; 2022:140-150.
138. Zhang Y, He N, Yang J, et al. mmformer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer; 2022:107-117.
139. Lin J, Lin J, Lu C, et al. CKD-TransBTS: clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *IEEE Trans Med Imag*. 2023;42(8):1-2461. https://doi.org/10.1109/TMI.2023.3250474
140. Gai D, Zhang J, Xiao Y, Min W, Zhong Y, Zhong Y. RMTF-Net: residual mix transformer fusion net for 2D brain tumor segmentation. *Brain Sci*. 2022;12(9):1145. https://doi.org/10.3390/brainsci12091145
141. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf Fusion*. 2023;91:376-387. https://doi.org/10.1016/j.inffus.2022.10.022
142. Nian R, Zhang G, Sui Y, et al. 3D brainformer: 3D fusion transformer for brain tumor segmentation. 2023:arXiv:2304.14508. https://doi.org/10.48550/arXiv.2304.14508
143. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600-612. https://doi.org/10.1109/TIP.2003.819861
144. Yin Z, Zhang Q, Zhao Y, et al. Prevalence and procedural risk of intracranial atherosclerotic stenosis coexisting with unruptured intracranial aneurysm. *Stroke*. 2023;54(6):1484-1493. https://doi.org/10.1161/STROKEAHA.122.041553
145. Chen C, Homma A, Mok VCT, et al. Alzheimer's disease with cerebrovascular disease: current status in the Asia–Pacific region. *J Intern Med*. 2016;280(4):359-374. https://doi.org/10.1111/joim.12495
146. Taher F, Mahmoud A, Shalaby A, El-Baz A. A review on the cerebrovascular segmentation methods. In: *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE; 2018:359-364.
147. Chen C, Zhou K, Wang Z, Xiao R. Generative consistency for semi-supervised cerebrovascular segmentation from TOF-MRA. *IEEE Trans Med Imag*. 2023;42(2):346-353. https://doi.org/10.1109/tmi.2022.3184675
148. Wu Q, Chen Y, Huang N, Yue X. Weakly-supervised cerebrovascular segmentation network with shape prior and model indicator. In: *Proceedings of the 2022 International Conference on Multimedia Retrieval*; 2022:668-676.
149. Li Y, Zhang Q, Zhou H, Li J, Li X, Li A. Cerebrovascular segmentation from mesoscopic optical images using Swin Transformer. *J Innov Opt Health Sci*. 2023;16(04):2350009. https://doi.org/10.1142/S1793545823500098
150. Gong H, Xu D, Yuan J, et al. High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nat Commun*. 2016;7(1):12142. https://doi.org/10.1038/ncomms12142
151. Zhong Q, Li A, Jin R, et al. High-definition imaging using line-illumination modulation microscopy. *Nat Methods*. 2021;18(3):309-315. https://doi.org/10.1038/s41592-021-01074-x
152. Rao VM, Wan Z, Arabshahi S, et al. Improving across-dataset brain tissue segmentation using transformer. *Front Neuroimaging*. 2022;1:1023481. https://doi.org/10.3389/fnimg.2022.1023481
153. Zhang J, Zhao L, Zeng J, Qin P. SF-SegFormer: stepped-fusion segmentation transformer for brain tissue image via inter-group correlation and enhanced multi-layer perceptron. In: *Medical image understanding and analysis*. Springer; 2022:508-518.
154. Sun Q, Fang N, Liu Z, Zhao L, Wen Y, Lin H. HybridCTrm: bridging CNN and transformer for multimodal brain image segmentation. *J Healthc Eng*. 2021;2021:7467261. https://doi.org/10.1155/2021/7467261
155. Yu X, Yang Q, Zhou Y, et al. UNesT: local spatial representation learning with hierarchical transformer for efficient medical segmentation. 2022:arXiv:2209.14378. https://doi.org/10.48550/arXiv.2209.14378
156. Zhang S, Ren B, Yu Z, et al. TW-Net: transformer weighted network for neonatal brain MRI segmentation. *IEEE J Biomed Health Inform*. 2023;27(2):1072-1083. https://doi.org/10.1109/JBHI.2022.3225475
157. Zhang H, Zhang J, Wang R, et al. Geometric loss for deep multiple sclerosis lesion segmentation. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2021:24-28.
158. Clough JR, Byrne N, Oksuz I, Zimmer VA, Schnabel JA, King AP. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Trans Pattern Anal Mach Intell*. 2020;44(12):8766-8778. https://doi.org/10.1109/TPAMI.2020.3013679
159. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*. 2011;54(3):2033-2044. https://doi.org/10.1016/j.neuroimage.2010.09.025
160. Rohé M-M, Datar M, Heimann T, Sermesant M, Pennec X. SVF-Net: learning deformable image registration using shape matching. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*. Springer; 2017:266-274.
161. Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: fast predictive image registration – a deep learning approach. *NeuroImage*. 2017;158:378-396. https://doi.org/10.1016/j.neuroimage.2017.07.008
162. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Trans Med Imag*. 2019;38(8):1788-1800. https://doi.org/10.1109/TMI.2019.2897538
163. Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Med Image Anal*. 2019;57:226-236. https://doi.org/10.1016/j.media.2019.07.006
164. Li Y, Zhang X, Chen D. Csrnet: dilated convolutional neural networks for understanding the highly congested scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2018:1091-1100.
165. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. 2017:arXiv:1706.05587. https://doi.org/10.48550/arXiv.1706.05587
166. Chen J, Frey EC, He Y, Segars WP, Li Y, Du Y. Transmorph: transformer for unsupervised medical image registration. *Med Image Anal*. 2022;82:102615. https://doi.org/10.1016/j.media.2022.102615
167. Chen J, He Y, Frey EC, Li Y, Du Y. Vit-v-net: vision transformer for unsupervised volumetric medical image registration. 2021:arXiv preprint arXiv:2104.06468. https://doi.org/10.48550/arXiv.2104.06468
168. Lee MC, Oktay O, Schuh A, Schaap M, Glocker B. Image-and-spatial transformer networks for structure-guided image registration. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer; 2019:337-345.
169. Mok TC, Chung A. Affine medical image registration with coarse-to-fine vision transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022:20835-20844.
170. Yang T, Bai X, Cui X, Gong Y, Li L. GraformerDIR: graph convolution transformer for deformable image registration. *Comput Biol Med*. 2022;147:105799. https://doi.org/10.1016/j.compbiomed.2022.105799

171. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. 2016:arXiv:1609.02907. https://doi.org/10.48550/arXiv.1609.02907
172. Ma M, Xu Y, Song L, Liu G. Symmetric transformer-based network for unsupervised image registration. *Knowl Base Syst*. 2022;257:109959. https://doi.org/10.1016/j.knosys.2022.109959
173. Zhang Y, Pei Y, Zha H. Learning dual transformer network for diffeomorphic registration. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer; 2021:129-138.
174. Zhu Y, Lu S. Swin-voxelmorph: a symmetric unsupervised learning model for deformable medical image registration using swin transformer. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer; 2022:78-87.
175. Qureshi MNI, Oh J, Lee B. 3D-CNN based discrimination of schizophrenia using resting-state fMRI. *Artif Intell Med*. 2019;98:10-17. https://doi.org/10.1016/j.artmed.2019.06.003
176. Huang H, Hu X, Zhao Y, et al. Modeling task fMRI data via deep convolutional autoencoder. *IEEE Trans Med Imag*. 2018;37(7):1551-1561. https://doi.org/10.1109/TMI.2017.2715285
177. Asadi N, Olson IR, Obradovic Z. A transformer model for learning spatiotemporal contextual representation in fMRI data. *Netw Neurosci*. 2023;7(1):22-47. https://doi.org/10.1162/netn_a_00281
178. Malkiel I, Rosenman G, Wolf L, Hendler T. Pre-training and fine-tuning transformers for fmri prediction tasks. 2021:arXiv preprint arXiv:211205761.
179. Deng X, Zhang J, Liu R, Liu K. Classifying ASD based on time-series fMRI using spatial–temporal transformer. *Comput Biol Med*. 2022;151:106320. https://doi.org/10.1016/j.compbiomed.2022.106320
180. Nguyen S, Ng B, Kaplan AD, Ray P. Attend and decode: 4D fmri task state decoding using attention models. In: *Proceedings of the Machine Learning for Health NeurIPS Workshop*. PMLR; 2020:267-279.
181. Dai W, Zhang Z, Tian L, et al. BrainFormer: a hybrid CNN-transformer model for brain fMRI data classification. 2022:arXiv:2208.03028. https://doi.org/10.48550/arXiv.2208.03028
182. Bedel HA, Şıvgın I, Dalmaz O, Dar SUH, Çukur T. BolT: fused window transformers for fMRI time series analysis. arXiv preprint arXiv:220511578. 2022. https://doi.org/10.1016/j.media.2023.102841
183. Li W, Wang S, Liu G. Transformer-based model for fMRI data: ABIDE results. In: *2022 7th International Conference on Computer and Communication Systems (ICCCS)*. IEEE; 2022:162-167.
184. Yu X, Zhang L, Zhao L, Lyu Y, Liu T, Zhu D. Disentangling spatial-temporal functional brain networks via twin-transformers. 2022:arXiv:2204.09225. https://doi.org/10.48550/arXiv.2204.09225
185. Zhao L, Wu Z, Dai H, et al. Embedding human brain function via transformer. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer; 2022:366-375.
186. Kumar S, Sumers TR, Yamakoshi T, et al. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*. 2022:2022-06. https://doi.org/10.1101/2022.06.08.495348
187. Hu J, Huang Y, Wang N, Dong S. BrainNPT: pre-training of Transformer networks for brain network classification. 2023:arXiv:2305.01666. https://doi.org/10.48550/arXiv.2305.01666
188. Sarraf S, Sarraf A, DeSouza DD, Anderson JA, Kabia M, Initiative AsDN. OViTAD: optimized vision transformer to predict various stages of Alzheimer's disease using resting-state fMRI and structural MRI data. *Brain Sci*. 2023;13(2):260. https://doi.org/10.3390/brainsci13020260
189. Kan X, Dai W, Cui H, Zhang Z, Guo Y, Yang C. Brain network transformer. In: *Advances in Neural Information Processing Systems*; 2022:25586-25599.
190. Liu H, Yang YI, Wang Y, et al. Ketogenic diet for treatment of intractable epilepsy in adults: a meta-analysis of observational studies. *Epilepsia Open*. 2018;3(1):9-17. https://doi.org/10.1002/epi4.12098
191. Ruffini G, Ibañez D, Castellano M, Dunne S, Soria-Frisch A. EEG-driven RNN classification for prognosis of neurodegeneration in at-risk patients. In: *Artificial Neural Networks and Machine Learning – ICANN 2016*. Springer International Publishing; 2016:306-313.
192. Supakar R, Satvaya P, Chakrabarti P. A deep learning based model using RNN-LSTM for the detection of schizophrenia from EEG data. *Comput Biol Med*. 2022;151:106225. https://doi.org/10.1016/j.compbiomed.2022.106225
193. Wan Z, Li M, Liu S, Huang J, Tan H, Duan W. EEGformer: a transformer–based brain activity classification method using EEG signal. *Front Neurosci*. 2023;17:1148855. https://doi.org/10.3389/fnins.2023.1148855
194. Xie J, Zhang J, Sun J, et al. A Transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification. *IEEE Trans Neural Syst Rehabil Eng*. 2022;30:2126-2136. https://doi.org/10.1109/TNSRE.2022.3194600
195. Song Y, Jia X, Yang L, Xie L. Transformer-based spatial-temporal feature learning for EEG decoding. 2021:arXiv:2106.11170. https://doi.org/10.48550/arXiv.2106.11170
196. Lee Y-E, Lee S-H. EEG-transformer: self-attention from transformer architecture for decoding EEG of imagined speech. In: *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE; 2022:1-4.
197. Lee S-H, Lee M, Lee S-W. Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication. *IEEE Trans Neural Syst Rehabil Eng*. 2020;28(12):2647-2659. https://doi.org/10.1109/TNSRE.2020.3040289
198. Liu J, Zhang L, Wu H, Zhao H. Transformers for EEG emotion recognition. 2021:arXiv:2110.06553. https://doi.org/10.1109/JSEN.2022.3144317
199. Du Y, Xu Y, Wang X, Liu L, Ma P. EEG temporal–spatial transformer for person identification. *Sci Rep*. 2022;12(1):14378. https://doi.org/10.1038/s41598-022-18502-3
200. Wang Z, Wang Y, Hu C, Yin Z, Song Y. Transformers for EEG-based emotion recognition: a hierarchical spatial information learning model. *IEEE Sensor J*. 2022;22(5):4359-4368. https://doi.org/10.1109/JSEN.2022.3144317
201. Kostas D, Aroca-Ouellette S, Rudzicz F. BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Front Hum Neurosci*. 2021;15:653659. https://doi.org/10.3389/fnhum.2021.653659
202. Tao Y, Sun T, Muhamed A, et al. Gated transformer for decoding human brain EEG signals. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE; 2021:125-130.
203. Siddhad G, Gupta A, Dogra DP, Roy PP. Efficacy of transformer networks for classification of raw EEG data. 2022:arXiv:2202.05170. https://doi.org/10.48550/arXiv.2202.05170
204. Ahn H-J, Lee D-H, Jeong J-H, Lee S-W. Multiscale convolutional transformer for EEG classification of mental imagery in different modalities. *IEEE Trans Neural Syst Rehabil Eng*. 2023;31:646-656. https://doi.org/10.1109/TNSRE.2022.3229330
205. Ma Y, Song Y, Gao F. A novel hybrid CNN-Transformer model for EEG Motor Imagery classification. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE; 2022:1-8.
206. Kroupi E, Vesin J-M, Ebrahimi T. Subject-independent odor pleasantness classification using brain and peripheral signals. *IEEE Trans Affect Comput*. 2015;7(4):422-434. https://doi.org/10.1109/TAFFC.2015.2496310
207. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(1):1929-1958.
208. Sun J, Xie J, Zhou H. EEG classification with transformer-based models. In: *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE; 2021:92-93.

209. Cho H, Ahn M, Ahn S, Kwon M, Jun SC. EEG datasets for motor imagery brain–computer interface. *GigaScience*. 2017;6(7):gix034. https://doi.org/10.1093/gigascience/gix034
210. Pfurtscheller G, Brunner C, Schlögl A, Da Silva FL. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage*. 2006;31(1):153-159. https://doi.org/10.1016/j.neuroimage.2005.12.003
211. Baevski A, Mohamed A. Effectiveness of self-supervised pre-training for ASR. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2020:7694-7698.
212. Jun E, Jeong S, Heo D-W, Suk H-I. Medical transformer: universal brain encoder for 3D MRI analysis. 2021:arXiv:2104.13633. https://doi.org/10.48550/arXiv.2104.13633
213. Li C, Cui Y, Luo N, et al. Trans-ResNet: integrating transformers and CNNs for Alzheimer's disease classification. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2022:1-5.
214. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. In: *Advances in Neural Information Processing Systems*; 2022:27730-27744.
215. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*; 2020:1877-1901.
216. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
217. Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. ChatCAD: interactive computer-aided diagnosis on medical image using large language models. 2023:arXiv:2302.07257. https://doi.org/10.48550/arXiv.2302.07257
218. Kirillov A, Mintun E, Ravi N, et al. Segment anything. 2023:arXiv:2304.02643. https://doi.org/10.48550/arXiv.2304.02643
219. Ma J, Wang B. Segment anything in medical images. 2023:arXiv:2304.12306. https://doi.org/10.48550/arXiv.2304.12306
220. Zhang C, Xu Y, Shen Y. Compconv: a compact convolution module for efficient feature learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*; 2021:3012-3021.
221. Chen C, Zhou K, Qi S, Lu T, Xiao R. A learnable Gabor Convolution kernel for vessel segmentation. *Comput Biol Med*. 2023;158:106892. https://doi.org/10.1016/j.compbiomed.2023.106892
222. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014:arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980
223. Lin T-Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; 2017.
224. Ma J, Chen J, Ng M, et al. Loss odyssey in medical image segmentation. *Med Image Anal*. 2021;71:102035. https://doi.org/10.1016/j.media.2021.102035
225. Li Y, Mao H, Girshick R, He K. *Exploring Plain Vision Transformer Backbones for Object Detection*. Springer Nature Switzerland; 2022:280-296.
226. Wang S, Gao J, Li Z, Zhang X, Hu W. A closer look at self-supervised lightweight vision transformers. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR; 2023:35624-35641.
227. Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020:10076-10085.
228. Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: training vision transformers from scratch on imagenet. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021:558-567.
229. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv*. 2023;55(9):1-35. https://doi.org/10.1145/3560815
230. Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. 2020:arXiv:2012.15723. https://doi.org/10.48550/arXiv.2012.15723
231. Zheng Z, Yue X, Wang K, You Y. Prompt vision transformer for domain generalization. 2022:arXiv:2208.08914. https://doi.org/10.48550/arXiv.2208.08914
232. Saeed N, Ridzuan M, Al Majzoub R, Yaqub M. Prompt-based tuning of transformer models for multi-center medical image segmentation. 2023:arXiv:2305.18948. https://doi.org/10.48550/arXiv.2305.18948
233. Chen H, Wang Y, Guo T, et al. Pre-trained image processing transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2021:12299-12310.
234. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022:16000-16009.
235. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. 2020:arXiv:2004.05150. https://doi.org/10.48550/arXiv.2004.05150
236. Shaharabany T, Dahan A, Giryes R, Wolf L. AutoSAM: adapting SAM to medical images by overloading the prompt encoder. 2023:arXiv:2306.06370. https://doi.org/10.48550/arXiv.2306.06370
237. LaBella D, Adewole M, Alonso-Basanta M, et al. The ASNR-MICCAI brain tumor segmentation (BraTS) challenge 2023: intracranial meningioma. 2023:arXiv:2305.07642. https://doi.org/10.48550/arXiv.2305.07642
238. Kossen T, Subramaniam P, Madai VI, et al. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Comput Biol Med*. 2021;131:104254. https://doi.org/10.1016/j.compbiomed.2021.104254

How to cite this article: Chen C, Wang H, Chen Y, et al. Understanding the brain with attention: a survey of transformers in brain sciences. *Brain-X*. 2023;1:e29. https://doi.org/10.1002/brx2.29