DenseFormer-MoE: A Dense Transformer Foundation Model with Mixture of Experts for Multi-Task Brain Image Analysis

Rizhi Ding, Hui Lu, Manhua Liu, *Member, IEEE*

***Abstract*— Deep learning models have been widely investigated for computing and analyzing brain images across various downstream tasks such as disease diagnosis and age regression. Most existing models are tailored for specific tasks and diseases, posing a challenge in developing a foundation model for diverse tasks. This paper proposes a Dense Transformer Foundation Model with Mixture of Experts (DenseFormer-MoE), which integrates dense convolutional network, Vision Transformer and Mixture of Experts (MoE) to progressively learn and consolidate local and global features from T1-weighted magnetic resonance images (sMRI) for multiple tasks including diagnosing multiple brain diseases and predicting brain age. First, a foundation model is built by combining the vision Transformer with Densenet, which are pre-trained with Masked Autoencoder and self-supervised learning to enhance the generalization of feature representations. Then, to mitigate optimization conflicts in multi-task learning, MoE is designed to dynamically select the most appropriate experts for each task. Finally, our method is evaluated on multiple renowned brain imaging datasets including UK Biobank (UKB), Alzheimer's Disease Neuroimaging Initiative (ADNI), and Parkinson's Progression Markers Initiative (PPMI). Experimental results and comparison demonstrate that our method achieves promising performances for prediction of brain age and diagnosis of brain diseases.**

***Index Terms*— Foundation Model, Mixture of Experts, Self-Supervised Learning, Multi-Task Learning, Transformer, Brain Disease.**

I. Introduction

BRAIN images such as magnetic resonance images (MRIs) have provided crucial insights into the structure and function of human brain [1], [2]. Due to their powerful capability in capturing disease-induced brain atrophy under non-invasive or minimally invasive conditions, these images are widely used in the clinical diagnosis of brain disorders such as Alzheimer's disease (AD) and Parkinson's disease (PD) [3], [4]. However, analyzing these intricate 3D images requires sophisticated techniques capable of extracting meaningful information from vast datasets. Deep learning, renowned for its ability to learn hierarchical representations of data, has emerged as a promising method for brain imaging analysis.

Rizhi Ding, and Manhua Liu are with the MoE Laboratory of Artificial-Intelligence, AI Institute, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai200240,China

Hui Lu is with the SJTU-Yale Joint Center of Biostatistics and Data Science, National Center for Translational Medicine, Shanghai Jiao Tong University, 800 Dongchuan Rd, Shanghai, 200240, Shanghai, China

In recent years, deep learning models have garnered extensive research attention for image computation and analysis, providing powerful tools for automating the detection, diagnosis, and prognosis of various brain diseases [1]–[6]. Deep convolutional neural networks (CNNs) have achieved remarkable success in various tasks like AD diagnosis and brain age prediction [1], [5], [7], [8]. These deep models leverage large datasets and sophisticated algorithms to detect patterns and anomalies in brain images that were previously unattainable with traditional methods.

However, most of these deep models are tailored for specific single tasks, significantly hindering their broader applicability [1], [2], [4], [9]. Despite their prowess in specialized domains, these models fail to offer a comprehensive solution for the diverse and multifaceted challenges posed by brain imaging tasks. For instance, the deep learning model optimized for detecting Parkinson's disease may not be effective for diagnosing other neurodegenerative diseases or predicting brain age in different contexts [4]. This lack of generalization limits their practicality in clinical settings, where a wide range of diagnostic and predictive tasks are necessary.

Moreover, these models typically rely on single architectural framework, either purely based on CNNs [8], [10] or Transformers [6], [11]. CNN-based models excel at local feature extraction but often struggle with capturing long-range dependencies. In contrast, Transformer-based models are adept at capturing long-range correlation and global context but may not perform as well in fine-grained feature extraction. This dependence on single-architecture designs poses significant limitations, as comprehensive brain health assessments typically require both global and local information for accurate diagnosis and prediction in various conditions. Therefore, with the growing availability of large-scale brain imaging datasets, there is a critical need for a foundation model that can robustly handle diverse brain imaging tasks, providing a unified and efficient approach to multi-task brain image analysis.

To address the above issues, this paper proposes a dense transformer foundational model with Mixture of Experts (DenseFormer-MoE) for multi-task brain image analysis, encompassing brain age prediction and multi-disease diagnosis. First, a DenseFormer backbone is constructed by combining the densely connected convolution neural network (Densenet) [5] with the self-attention Transformer [6]. This integration enhances the model's capacity to capture both local and global feature representation. Second, to foster generalization

of the backbone network, the self-supervised pre-training is conducted using the Masked Autoencoder (MAE) [12], which reconstruct the missing sections by masking a substantial portion of input data to learn robust and diverse feature representations from large-scale, unlabeled data. Finally, the Mixture of Experts (MoE) [13] is integrated with the backbone to dynamically select appropriate experts for different tasks, thus improving performance in complex multi-task scenarios. The proposed method aims to provide a comprehensive solution for brain imaging analysis, enabling concurrent execution of multiple diagnostic tasks while enhancing its efficiency and effectiveness. Different from the existing methods, the main contributions of this paper are summarized as follows:

- **Hybrid Backbone Network:** The proposed DenseFormer backbone network integrates the dense connectivity of Densenet with the Transformer's capability to capture the long-range dependency, thereby enhancing the model's ability to learn both local and global features for more comprehensive analysis of brain images.
- **Effective Self-Supervised Learning:** The Masked Autoencoder (MAE) is employed to pre-train the DenseFormer backbone through self-supervised learning, enabling it to extract rich and powerful features from large-scale unlabeled data. This methodology enhances the robustness and diversity of features, which is particularly advantageous in the medical brain imaging domain where the labeled data is scarce.
- **Dynamic Multi-Task Learning:** The proposed foundational model employs the MoE to dynamically select the most appropriate experts for various tasks. This strategy enhances the model's performance in multi-task learning scenarios by mitigating conflicts and interference among tasks, ensuring that each task benefits from specialized sub-models (experts).
- **Comprehensive Performance and Diagnostic Support:** The proposed DenseFormer-MoE model has been evaluated on multiple brain image datasets, demonstrating superior performance in various diagnostic tasks, such as prediction of brain age and diagnosis of multiple diseases, when compared to the existing state-of-the-art methods.

II. Related Work

This section briefly reviews the related works, focusing on the methods of deep feature learning, self-supervised learning and multi-task learning in brain image analysis.

A. Deep Feature Learning in Brain Image Analysis

In recent years, deep learning models have been widely investigated for brain image analysis, improving the accuracy and efficiency of various diagnostic tasks such as brain disease diagnosis and age estimation. Convolutional neural networks (CNNs) and Transformer, in particular, have emerged as two powerful deep learning architectures for feature extraction and classification in a wide range of applications, including brain image segmentation, disease diagnosis, and anatomical structure analysis [8], [14].

CNNs have been extensively employed in brain image analysis due to their ability to learn hierarchical features from imaging data, and have shown remarkable accuracy in detecting and classifying brain diseases such as Alzheimer's and Parkinson's diseases [4], [8], [14]. Early CNN architectures such as LeNet-5 [15] cascade multiple convolutional and pooling layers. Recent advancements in architecture including ResNet [7] and Densenet [5] have further improved the performance of CNNs in brain image analysis. They distinguish brain diseases like AD and PD from healthy controls by analyzing structural MRI scans, benefiting from deep architectures that extract complex patterns associated with these conditions. CNNs can automatically extract discriminative features from images without manual feature engineering, which facilitate learning the complex patterns and features that are useful for brain image analysis in clinical practice. However, CNNs also have some limitations in brain imaging analysis. For example, they may suffer from overfitting when trained on small datasets. Additionally, CNNs tend to focus on local features and may miss global contextual information, which is important for tasks such as anatomical structure analysis.

Transformers, originally developed for natural language processing, have recently gained attention in the field of computer vision due to their ability to model long-range dependencies and capture global context [6]. Unlike CNNs, which process images locally, Transformers utilize self-attention mechanisms to learn long-range correlations to extract comprehensive and contextualized features. Vision Transformer (ViT) is a notable architecture that divides images into fixed-size patches and treating each patch as a token [16]. These tokens are processed by a Transformer encoder to learn relationships between patches. A global-local transformer model integrates global context and local details from MRI scans using an attention mechanism, achieving high accuracy in brain age estimation [11]. However, such methods have limitation in learning fine-grained local features, crucial for detailed spatial tasks.

To address these limitations, recent approaches combined CNNs and Transformers to leverage the strengths of both architectures [17], [18]. By integrating the local feature extraction of CNNs with the global context modeling of Transformers, these hybrid models achieve more comprehensive feature learning. For instance, the CoTr model employs a CNN to extract feature representations and a Transformer with deformable self-attention to efficiently model long-range dependencies [18]. However, these approaches lacks effective interaction between CNN and Transformer components.

B. Self-Supervised Learning in Brain Image Analysis

Supervised learning has been the dominant approach in brain image analysis, relying on large amounts of labeled data, which are often scarce and expensive in the medical field. Self-supervised learning (SSL) offers a promising alternative by enabling models to learn from unlabeled data through pretext tasks that generate supervisory signals. This allows for pre-training on vast amounts of unlabeled data and fine-tuning on smaller labeled datasets for specific medical tasks.

A common approach in self-supervised learning is reconstruction-based methods, such as Convolutional Autoen-

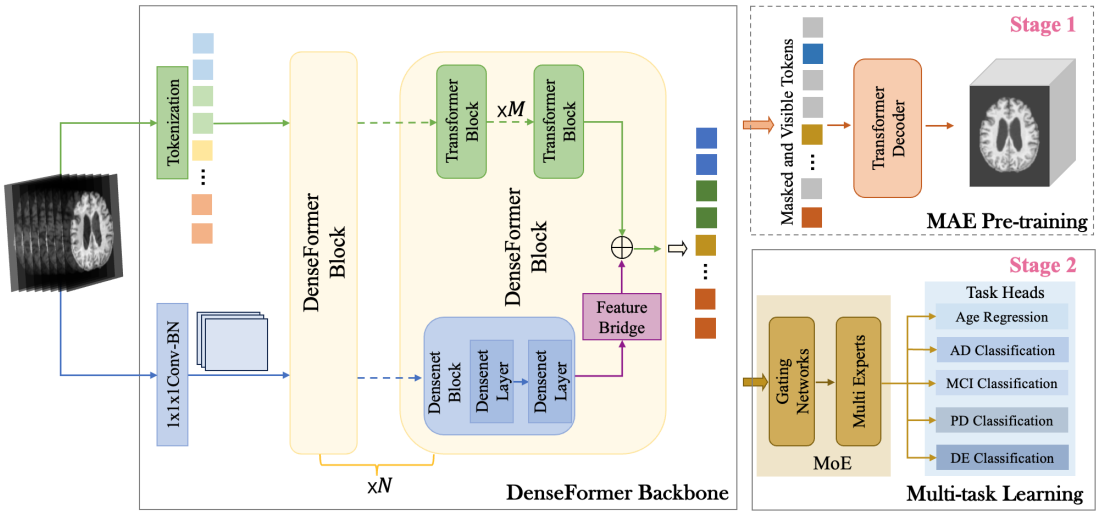


Fig. 1. Overview of the dense transformer foundational model with Mixture of Experts (DenseFormer-MoE) for multi-task brain image analysis, which consists of a DenseFormer backbone by integration of Densenet and Transformer, self-pre-training by MAE and multi-task learning by MoE.

coders (CAE) and Variational Autoencoders (VAE) [17], [19]. CAE use CNN layers to downsample and learn features from the input image, then upsample and reconstruct the image. This helps capture important structural features beneficial for downstream tasks. A two-stream convolutional autoencoder was proposed to learn latent representations from multimodal MRI data by reconstructing input images [17]. Variational Autoencoders (VAE) learn to encode input images into a latent space and then decode them back to the original images [19]. This probabilistic framework not only reconstructs images but also generates new samples, providing robust feature representations useful for various tasks. However, CAE can struggle with capturing global context due to their localized feature learning, and VAE often produce blurry reconstructions and less sharp feature extraction due to their generative nature.

Recently, another reconstruction-based self-supervised learning method, Masked Autoencoders (MAE), was proposed for capturing both detailed local features and broader contextual information [12]. MAE masked a portion of the input image and train the model to reconstruct the missing parts. This method forces the model to understand both the local and global context, leading to richer feature representations. Such comprehensive feature learning benefits various downstream tasks, including disease diagnosis and age estimation. One study successfully applied MAE to medical image classification and segmentation, such as chest X-ray disease classification and MRI brain tumor segmentation [20]. Extending the application of MAE to brain image analysis could unlock new potential in accurately diagnosing and understanding various neurological conditions.

Additionally, contrastive learning has been widely investigated as a powerful self-supervised learning technique in medical image analysis to learn feature representations such that similar instances are close together in the representation space while dissimilar instances are far apart [21]–[26]. It enables models to capture relevant features by learning the similarities from large amounts of unlabeled data and improve the performance even with limited labeled data. Contrastive learning methods such as SimCLR [23], MoCo [24], SimSiam [25] and DINO [26] have further advanced the field by proposing effective frameworks for representation learning. For instance, MoCo introduces a momentum encoder to maintain a large and consistent dictionary for contrastive learning, improving feature quality for downstream tasks. SimSiam eliminates the need for negative pairs by using stop-gradient techniques, simplifying the contrastive learning framework while maintaining competitive performance. SimCLR, on the other hand, uses a simple yet effective approach by employing large-batch training and a contrastive loss function, achieving significant performance improvements in unsupervised learning tasks. Contrastive learning is applied for brain MRI segmentation, showing significant improvements by pre-training on unlabeled data and fine-tuning on smaller labeled sets [21]. Contrastive pre-training is used for pathology image classification, achieving substantial gains in cancer detection tasks [22]. By leveraging large-scale unlabeled medical data, contrastive learning facilitates meaningful and transferable feature extraction for medical image analysis.

C. Multi-Task Learning in Brain Image Analysis

To achieve multi-task medical diagnosis, a traditional approach is to train a dedicated model for each task, which makes training and deployment cumbersome and inefficient. A common solution is multi-task learning (MTL), where a single end-to-end model performs multiple tasks. This method efficiently utilizes data, reduces computational resources, and enhances the model's generalization. In brain image analysis, MTL can concurrently handle disease diagnosis, segmentation, and age estimation, resulting in improved diagnostic accuracy and efficiency compared to single-task models.

The traditional MTL model trains shared CNNs to capture underlying patterns beneficial for multiple tasks, forming a strong basis for task-specific layers [27]. For instance, a multi-task CNN architecture was used for segmenting six brain tissues, pectoral muscle in breast MRI, and coronary arteries

in cardiac CTA. The loss function of MTL combines all task losses through a weighted sum, allowing the model to optimize for all tasks simultaneously [27].

A multi-task learning approach was proposed to simultaneously predict multiple diseases by analyzing correlations within electronic medical records, thereby improving diagnosis performance by leveraging disease interdependencies [28]. Chelaramani et al. developed an MTL model integrating knowledge distillation to diagnose eye diseases from retinal images. This approach simultaneously classifies broad disease categories, sub-categories, and generates detailed diagnoses with high accuracy, especially valuable in small-data scenarios [29]. In Parkinson's disease prediction, an MTL model was developed to predict disease risk by integrating neuroimaging data with patient demographics, achieving a high AUC and improving early diagnostic capabilities in Parkinson's disease [30]. However, shared layers may not always capture task-specific features effectively, leading to suboptimal performance in some tasks. Additionally, balancing the contributions of each task in the loss function can be challenging, often requiring careful tuning of the weights to ensure that no single task dominates the learning process.

To address these limitations, recent studies have introduced methods like GradNorm and Uncertainty Weighted Loss (UWL) [31], [32]. GradNorm dynamically adjusts the weights of each task based on their gradients, ensuring that tasks with slower convergence do not hinder overall learning [31]. This adaptive approach promotes balanced training and improves model performance across all tasks. On the other hand, UWL assigns weights to tasks based on the uncertainty of their predictions, allowing the model to focus more on tasks that are more uncertain, thus enhancing the robustness of the learning process [32]. More recently, the Mixture of Experts (MoE) framework has been proposed for MTL [13], [33], [34]. MoE dynamically selects a subset of expert networks for each task, allowing for task-specific learning while still benefiting from shared knowledge. This approach leverages specialized sub-models, or "experts," tailored to specific tasks, which can significantly improve model adaptability and performance. By dynamically assigning the most relevant experts to each task, MoE mitigates task interference and enhances overall accuracy and efficiency. For example, a modularized MoE framework was proposed to integrate the MoE layers into a vision transformer, ensuring that only a small set of experts is activated for each task [13]. Gao et al. [33] proposed a Task-Customized Mixture of Adapters (TCA) for general image fusion, which uses a similar MoE approach to select the most relevant adapters for each specific task, thereby improving task performance and resource efficiency. In brain image analysis, applying MoE to MTL can optimize the concurrent learning of related tasks, enhancing performance and generalization.

III. Proposed Method

In this section, we present the dense transformer foundation model with Mixture of Experts (DenseFormer-MoE) for multi-task brain image analysis, as shown in Fig. 1. The DenseFormer-MoE consists of three main components: DenseFormer backbone, MAE self-supervised pre-training and multi-task learning by MOE, detailed as below.

A. DenseFormer Backbone

Generally, there are two crucial types of features for brain image analysis: local features capturing the intricate details of specific regions, and global features, which encapsulate the overall structure and context. CNN and Transformer are two important network structures widely used for learning image features. CNNs hierarchically learn local features through convolutional operations, whereas Transformers leverage multi-head self-attention modules to capture long-range correlations and extract global contextual features. A hybrid model was proposed to take advantage of both structures to learn the global and local features for visual recognition [18]. However, this approach did not fully exploit the potential of both architectures, limiting its feature learning capabilities. In this paper, we propose a hybrid dense Transformer (DenseFormer) backbone which effectively integrates Densenet and vision Transformer (ViT) networks in parallel. By continuously interacting and coupling these two structures, DenseFormer leverages their complementary advantages, as illustrated in Fig. 1. The DenseFormer consists of N DenseFormer blocks, each containing DenseNet and ViT streams that capture local and global features of sMRI data, respectively. Fig. 2 shows the structure of a DenseFormer block, detailed as follows.

First, the Densenet block is constructed with two Densenet layers to capture local features of sMRI data. Each Densenet layer consists of a 1x1x1 and a 3x3x3 convolutional layera, batch normalization and an activation layer. Given an input 3D brain volume $\mathbf{X} \in \mathbb{R}^{H\times W\times D\times 1}$, where H, W and D represent the height, width and depth of the image, respectively, the feature map $\mathbf{X}^i_{\text{dense}}$ of the i-th Densenet layer is computed in Densenet block as:

$$\mathbf{X}^i_{\text{dense}} = \sigma(\text{BN}(\text{Conv}(\mathbf{X}^{i-1}_{\text{dense}} \oplus \cdots \oplus \mathbf{X}^1_{\text{dense}}))) \quad (1)$$

where $\oplus$ denotes the concatenation of outputs from all previous dense layers, and σ represents the activation function. The Densenet layers in a block are densely connected with each other by receiving the feature maps of all preceding layers as input. At the end of Densenet block, a transition layer is added for down-sampling the feature maps. The Densenet block excel in feature reuse and ensure efficient gradient propagation to learn the discriminative local features.

Second, in each DenseFormer block, there are M Transformer blocks cascaded to capture global context and long-range dependencies. Each Transformer block consists of LayerNorm, Multi-head Self-attention (MHSA) and Multilayer Perceptron (MLP) modules. MHSA module is used to learn complex spatial transforms and long-distance feature dependencies. A set of token vectors obtained by linear projection of fixed-size image patches or previous Transformer block are input to each Transformer block. To retain spatial information, a positional encoding is added to token vector before being fed into the Transformer block. The feature map $\mathbf{X}^j_{\text{trans}}$ of the j-th Transformer block is computed as:

$$\mathbf{X}^j_{\text{trans}} = \text{MLP}(\text{LayerNorm}(\text{MHSA}(\mathbf{X}^{j-1}_{\text{trans}}) + \mathbf{X}^{j-1}_{\text{trans}})) \quad (2)$$

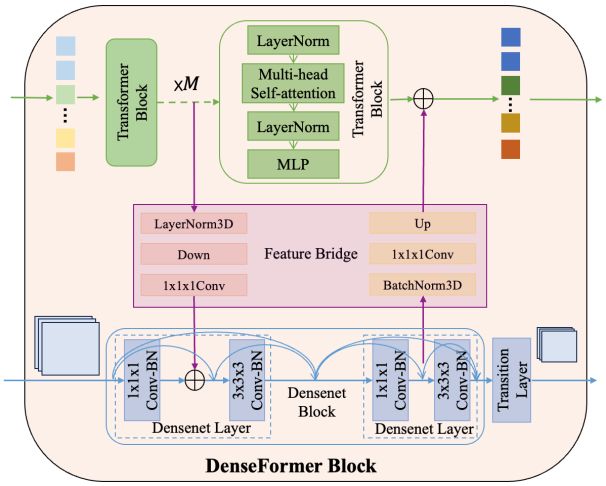


Fig. 2. Network structure of the DenseFormer block integrating Densenet block, Transformer blocks, and feature bridges to learn local and context features and their coupling, respectively.

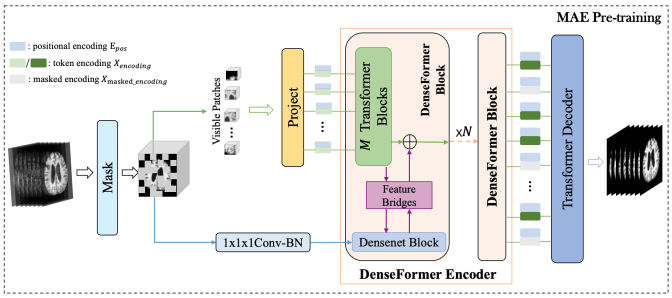


Fig. 3. The MAE self-supervised pre-training of DenseFormer backbone, consisting of patch partition and masking, projection, DenseFormer encoder and Transformer decoder for image reconstruction.

Finally, since the feature dimensions of Densenet and Transformer blocks are inconsistent, we design a feature bridge to align and fuse their feature maps with sparse cross-layer connectivity. The feature bridge consists of down-sampling, up-sampling, LayerNorm and BatchNorm modules. In the down-sampling module, tokens from the Transformer block are integrated into the feature maps of DenseNet layer by using LayerNorm3D, down-sampling and a 1x1x1 Conv3D layer. In the up-sampling module, the feature maps from Densenet layer are aligned and fused with the tokens from Transformer block by cascading BatchNorm3D, a 1x1x1 Conv3D layer for channel mapping, and up-sampling. There are one Densenet block and three Transformer blocks in each DenseFormer block. Thus, the Densenet block is aligned and fused with the last Transformer block by the feature bridge to combine the global and local features for more powerful representations.

To balance the tradeoff between the model capability and computation efficiency, the hybrid DenseFormer backbone comprises four DenseFormer blocks, with each block consisting of three Transformer blocks and one Densenet block. In the last DenseFormer block, the up-sampling bridge is used to convert the feature maps from Densenet block into token representations, which are fed to downstream tasks. The DenseFormer backbone facilitates the effective interaction and complementary fusion of local features from Densenet and global context features from Transformer, which can enhance feature representation for more accurate brain image analysis.

B. Self-Supervised Pre-training By MAE

After constructing the DenseFormer backbone, we employ the self-supervised pre-training by Masked Autoencoders (MAE) [12]. MAE aims to learn the feature representations by reconstructing missing or masked portions of input data, enabling DenseFormer to leverage large amounts of unlabeled data and enhance downstream task performance with limited labels. Fig. 3 shows the MAE self-supervised pre-training of DenseFormer backbone, which consists of patch partition and masking, DenseFormer encoder and Transformer decoder for image reconstruction, detailed as below.

First, the input MRI data $\mathbf{X} \in \mathbb{R}^{H\times W\times D\times 1}$ is uniformly partitioned into non-overlapping patches of size 8x8x8 and 75% of patches are randomly masked to obtain a new volume $\mathbf{X}_m$ (the masked values set to 0). In the Transformer stream, the visible patches are embedded by a channel projection layer into a sequence of tokens with added positional encodings, followed by a series of Transformer blocks for feature learning. For the Densenet stream, to prevent masked information from reaching the Transformer blocks, the masked volume $\mathbf{X}_m$ is input to a 3D convolutional layer, followed by 3D Densenet blocks for feature learning. Since the feature dimensions of Densenet and Transformer are inconsistent, Feature Bridge with upsampling and downsampling is used to align and fuse the feature maps from two streams. The DenseFormer encoder output concatenates features from both streams.

Then, the MAE decoder is a lightweight Transformer decoder as in [12], which consists of fewer layers than the encoder for efficient reconstruction of masked patches during training. Each decoder layer includes multi-head self-attention (MHSA), feedforward layers, and layer normalization. The decoder receives the full set of tokens arranged in their original patch order, including the visible tokens from the DenseFormer encoder and the masked tokens. The positional encodings are added to all tokens and each masked token is a learned vector indicating the presence of a missing patch to be predicted.

Finally, the DenseFormer backbone is pre-trained with the MAE self-supervised learning. By masking portions of the input patches and subsequently reconstructing the missing information, the model learns robust features without requiring supplementary data augmentation. The output of decoder is reshaped to form a reconstructed image. The MAE pre-training loss function is computed only on the masked patches, using the mean squared error (MSE) between the reconstructed patches and original ones as:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 \tag{3}$$

where $\mathcal{M}$ is the set of indices of masked patches, $\hat{\mathbf{x}}_i$ is the predicted patch at i, and $\mathbf{x}_i$ is its original patch.

It's worth noting that the decoder is used only for pre-training DenseFormer and is removed when the pre-trained

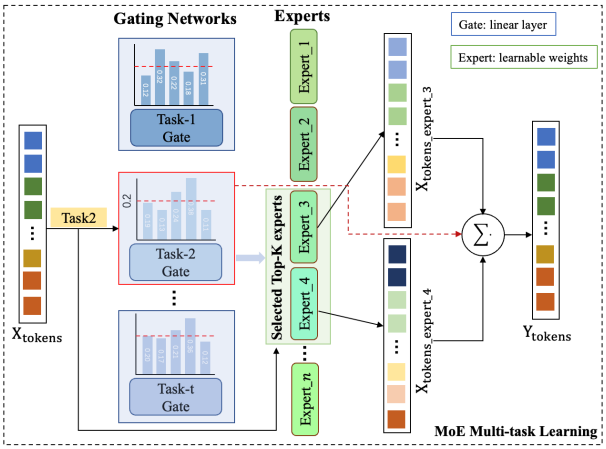


Fig. 4. The MoE structure for multi-task learning consisting of gating networks and expert networks.

model is applied to downstream tasks. Through MAE self-supervised learning, both the Densenet and Transformer streams in DenseFormer benefit from pre-training. This approach uses only partial patch information to reconstruct the original image, encouraging the network to capture local region correlations and uncover more powerful, representative features. Additionally, it leverages large-scale unlabeled sMRI data to extract meaningful, generalizable, and transferable features for various downstream tasks.

C. MoE for Multi-Task Learning

Multi-task learning aims to perform various tasks such as diagnosis of multiple diseases using a single model. Traditionally, a shared backbone network is trained for multiple tasks by optimizing a weighted sum of task-specific losses. However, this method does not adequately cater to the unique demands of each task. Especially for unrelated tasks, balancing feature learning becomes challenging, potentially resulting in suboptimal performance. To overcome these limitations, we integrate the MoE module into the DenseFormer backbone for multi-task learning. The MoE module facilitates the dynamic assignment of different experts to specific tasks, ensuring each task accesses the most relevant features. Fig. 4 illustrates the structure of MoE for multi-task learning, comprising the gating networks and multiple expert networks, as detailed below.

First, the gating network of MoE selects the most relevant experts based on the specific task input. It consists of a series of linear mapping layers with the softmax function to calculate the selection probability for each expert. Given the input feature tokens $\mathbf{X}_{\text{tokens}}$ obtained from the DenseFormer backbone, the gating network sparsely selects the most relevant experts using a Noisy Top-K Gating network in [13]. Assuming an MoE layer has T gating networks, denoted as $\mathbb{G} = \{G_1, G_2, ..., G_T\}$, where T is the number of tasks, the gating function G_t is computed as:

$$G_t(\mathbf{X}_{\text{tokens}}) = \text{TopK}(\text{Softmax}(\mathbf{X}_{\text{tokens}}\theta_g + \mathcal{N}(0,1)\log(1+\exp(\mathbf{X}_{\text{tokens}}\theta_{noise}))), \quad (4)$$

where θ_g and θ_{noise} are the learnable parameters of the gating and noise, respectively. The addition of noise perturbs the gating function, making its output smoother and more diverse to better adapt to various inputs. Thus, the gating network responds to the task input and calculates the probability of being selected for each expert. Finally, the number of the selected experts is fixed at topk.

Second, the expert networks play a crucial role in the MoE by enabling the model to specialize in various tasks through distinct sub-networks. This structure ensures that each expert concentrates on learning task-specific features. The expert networks generate expert-level responses by passing through a set of independently learnable sub-models, denoted as $\mathbb{E} = \{E_1, E_2, ..., E_n\}$, where n represents the number of experts. Each expert can be characterized as a lightweight network of learnable parameters, with its response function $E_n(\cdot)$ performing computations tailored to the specific task based on the token feature $\mathbf{X}_{\text{tokens}}$. Therefore, based on the expert selection probabilities in the gating network, the output from each expert is weighted and summed to generate the MoE output $\mathbf{Y}_{t,\text{tokens}}$ as:.

$$\mathbf{Y}_{t,\text{tokens}} = \sum_{n=1}^{topk} G_{t,n}(\mathbf{X}_{\text{tokens}}) \cdot E_n(\mathbf{X}_{\text{tokens}}) \quad (5)$$

where $G_{t,n}(\cdot)$ denotes the gating function of the n-th expert for the t-th task, and $\mathbf{Y}_{t,\text{tokens}}$ represents the final MoE output for task t. It is important to note that only the topk experts, selected based on the gating networks, participate in feature response. This design enables the model to assign different experts to various tasks, thus promoting specialized feature learning. By directing the input through experts tailored to the specific needs of each task, the MoE framework enhances the specialized acquisition of features across tasks, consequently minimizing the interference caused by unrelated tasks.

The advantages of MoE lie in its ability to dynamically encode features for different tasks, enriching feature diversity through multiple experts. This approach enhances prediction accuracy by allowing each expert to focus on task-specific features, while also reducing conflicts in joint task learning. By integrating the MoE with DenseFormer backbone, we have developed a foundational model (DenseFormer-MoE), which can be used for various image analysis tasks such as diagnosing various brain diseases.

D. Loss Function and Implementation

The proposed foundational model consists of DenseFormer backbone, the MAE self-supervised pre-training and the MoE multi-task learning. The DenseFormer backbone combines the Densenet and Transformer structures to extract robust features in diverse brain imaging tasks. It is composed of four DenseFormer blocks. To train the foundational model, the MAE self-supervised learning is used to pre-train the DenseFormer backbone, followed by the MoE multi-task learning for various downstream tasks of brain image analysis.

First, the DenseFormer backbone is pre-trained using MAE self-supervised learning on the UKB dataset. The MAE mask ratio is 0.75, and reconstruction is guided by a masked MSE loss function. The AdamW optimizer is used with a learning rate of 1.5e-4 and a weight decay of 0.05. The learning rate

TABLE I
DEMOGRAPHIC INFORMATION OF UKB, ADNI AND PPMI DATASETS.

Dataset	Group	Number	Gender(M/F)	Age
UKB	NC	16458	8572/7886	46-81
ADNI1	NC	227	117/110	65-95
	AD	196	99/97	56-93
	MCI	194	118/76	50-92
ADNI2	NC	200	94/106	58-93
	AD	156	90/66	56-92
	MCI	158	84/74	58-93
PPMI	NC	312	176/136	38-83
	PD	312	188/124	30-81
ABIDE	HC	571	464/107	6-57
	ASD	528	472/56	7-64
OASIS3	NC	171	73/98	47-94
	AD	174	88/86	43-89
	FD	30	13/17	44-92

follows a cosine annealing schedule with an initial warm-up phase, decreasing to zero. MAE pre-training is conducted over 300 epochs with a batch size of 16.

After pre-training the backbone, we fine-tune the MoE subnetwork of foundational model by multi-task learning for various downstream tasks including prediction of brain age and diagnosis of different brain disease such as AD, mild cognitive impairment (MCI), frontotemporal dementia (FD) and PD. For task of brain age prediction, the mean squared error (MSE) loss function is used to train the MoE subnetwork. The MSE loss is computed between the predicted brain age and the ground-truth age as:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{6}$$

where y_i and $\hat{y}_i$ denote the ground truth age and predicted brain age, respectively; N is the total number of training samples. For the disease diagnosis tasks, the Binary Cross-Entropy (BCE) loss function is used to train the MoE subnetwork for disease classification and it is computed as:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)], \tag{7}$$

where y_i is the ground truth label (0 for healthy and 1 for the disease), and $\hat{y}_i$ is the predicted probability of disease. In the training of the MoE subnetwork, the initial learning rate is 0.001, and a layer-wise learning rate decay strategy is employed with a decay rate of 0.75 and stochastic DropPath with a probability of 10%. All training task data are loaded into one dataloader for multi-task learning. The AdamW optimizer is used to train for 70 epochs with a batch size of 32.

IV. EXPERIMENTS AND RESULTS

In this section, we conduct the extensive experiments to evaluate the proposed method and compare the results with those of existing methods in the literature.

A. Datasets and Preprocessing

The datasets used in our experiments are the sMRI data obtained from multiple public database including UKBiobank (UKB), Alzheimer's Disease Neuroimaging Initiative (ADNI), Parkinson's Disease Progression Markers Initiative(PPMI), Autism Brain Imaging Data Exchange (ABIDE) and Open Access Series of Imaging Studies - 3 (OASIS3). UKB (https://biobank.ctsu.ox.ac.uk/) is a large-scale biomedical database with detailed health and genetic data of 500,000 volunteers from the United Kingdom, including genomic, health records, lifestyle, environmental data, and MRI imaging. ADNI study (http://adni.loni.usc.ed/) collected clinical assessments, brain images and biomarker data including sMRI, DTI and other biomarkers, from thousands of participants with AD, MCI and normal controls (NC) across North America for AD dementia research. PPMI (https://www.ppmi-info.org/data/) collected the clinical, imaging and biosample data from participants and aimed to identify biomarkers of PD progression. ABIDE (https://fcon_1000.projects.nitrc.org/indi/abide/) is a multi-site dataset that collects MRI data to study Autism Spectrum Disorder (ASD). OASIS3 (https://www.oasis-brains.org/) is a longitudinal dataset with the MRI data, clinical assessments, and cognitive evaluations from participants with AD, frontotemporal dementia (FD) and NC.

Demographic information of studied subjects for these datasets is summarized in TABLE I. T1-weighted sMRI data are used in our experiments. The UKB, PPMI, and ABIDE datasets are split into training and testing sets in an 8:2 ratio. For the UKB dataset, there are 13,166 training subjects and 3,292 testing subjects. Since the UKB training set is large, it is used to pre-train the backbone. For the PPMI dataset, there are 500 training subjects (PD: 250, NC: 250) and 124 testing subjects (PD: 62, NC: 62). For the ABIDE dataset, there are 879 training subjects (ASD: 423, NC: 456) and 220 testing subjects (ASD: 105, NC: 115). For the ADNI dataset, we use ADNI1 as the training set and ADNI2 as the testing set, similar to the previous study. Additionally, the MCI subjects are divided into stable MCI (sMCI) and progressive MCI (pMCI) based on whether subjects converted to AD within 36 months, where sMCI remains MCI at all time points and pMCI progresses to AD. In particular, ADNI1 (sMCI: 111, pMCI: 83) and ADNI2 (sMCI: 90, pMCI: 68). For the OASIS3 dataset, we randomly split the NC subjects into 137 training and 34 testing samples and the AD subjects into 139 training and 35 testing samples in an 8:2 ratio, while the FD subjects are randomly split in a 6:4 ratio into 18 training and 12 testing samples with five-fold cross-validation.

The sMRI data from the UKB dataset are preprocessed with a quality control pipeline as in [35], available to all UKB-authorized researchers. The FMRIB Software Library (FSL, https://fsl.fmrib.ox.ac.uk/) is used for image preprocessing including reorienting images, reducing the field of view, skull stripping, isotropic resolution adjustment, and image registration. Brain extraction is performed with the Brain Extraction Tool (BET) [36], and images are registered to the 1 mm resolution MNI152 template using FMRIB's Nonlinear Image Registration Tool (FNIRT). All sMRI data from the ADNI,

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT BACKBONE NETWORKS FOR AD AND PD DIAGNOSIS AND AGE PREDICTION.

Backbone	#Params (M)	AD vs. NC				PD vs. NC				Age Prediction		
		ACC↑	SEN↑	SPEC↑	AUC↑	ACC↑	SEN↑	SPEC↑	AUC↑	MAE↓	RMSE↓	PCC↑
Resnet50-3D [7]	46.16	0.845	0.859	0.835	0.912	0.802	0.888	0.714	0.820	3.602	4.778	0.785
Densenet3D [5]	1.59	0.870	0.807	0.920	0.935	0.817	0.936	0.698	0.836	3.167	3.961	0.865
ViT [16]	85.42	0.691	0.621	0.745	0.721	0.778	0.809	0.746	0.822	3.381	4.208	0.888
DenseFormer-S	22.06	0.874	0.821	0.915	0.923	0.817	0.920	0.714	0.827	2.795	3.558	0.887
DenseFormer-M	48.84	0.879	0.825	0.925	0.925	0.825	0.936	0.714	0.836	2.758	3.476	0.886
DenseFormer	**86.25**	**0.887**	**0.833**	**0.930**	**0.937**	**0.833**	**0.968**	**0.701**	**0.847**	**2.693**	**3.430**	**0.891**

PPMI, ABIDE and OASIS3 datasets are preprocessed with the N4 algorithm for bias field correction. They are further processed using a pipeline similar to that of UKB dataset on MNI152 template space. All registered images are cropped to size of 162×192×162 by removing the uninformative zero-value pixels. To balance the tradeoff between computation cost and discriminative feature learning, the images are further resized to 80×96×80 as the inputs of network.

In our experiments, mean absolute error (MAE), root means squared error (RMSE) and Pearson correlation coefficient (PCC) are computed for quantitative evaluation of brain age estimation, while the Accuracy (ACC), Sensitivity (SEN), Specificity (SPEC), and Area Under the Curve (AUC) are computed for the diagnosis classification tasks.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT DENSEFORMER BLOCKS FOR AD DIAGNOSIS ON ADNI DATASET.

Blocks Num	Accuracy	Sensitivity	Specificity	AUC
1	0.845	0.705	0.903	0.893
2	0.868	0.804	0.912	0.916
3	0.876	0.826	0.915	0.928
4	0.887	0.833	0.930	0.937
5	0.889	0.835	0.926	0.934

TABLE IV
PERFORMANCE COMPARISON OF AD DIAGNOSIS WITH AND WITHOUT FEATURE BRIDGE ON ADNI DATASET.

Feature Bridge	Accuracy	Sensitivity	Specificity	AUC
✓	0.887	0.833	0.930	0.937
✗	0.876	0.782	0.923	0.922

B. Effectiveness of DenseFormer Backbone

In this experiment, we conduct ablation study to test the effectiveness of the DenseFormer network structure and its MAE Pre-training.

Effectiveness of DenseFormer Structure: Since the Denseformer backbone is a hybrid network by combining Densenet and Transformer blocks, we conduct experiments to test the effectiveness of DenseFormer structure. First, we compare the Denseformer with other popular deep networks including the Resnet-50-3D [7], Densenet3D [5] and ViT [16]. In addition, we compare the proposed Denseformer, featuring an encoding dimension of 768 and 12 attention heads, with its down-scaled versions: Denseformer-S and Denseformer-M. Specifically, Denseformer-S has an encoding dimension of 384 and 6 attention heads, whereas Denseformer-M has an encoding dimension of 576 and 9 attention heads. The results on the ADNI, PPMI, and UKB datasets, including a comparison of performance metrics and the parameter count ("Params") for each backbone, are shown in TABLE II. We can see that the proposed Denseformer consistently performs better than other deep networks for AD and PD diagnosis and age prediction tasks. Specifically, it achieves the highest ACC of 0.887 and AUC of 0.937 for AD diagnosis, the highest ACC of 0.833 and AUC of 0.847 for PD diagnosis, and the best MAE of 2.693 and PCC of 0.891 for age prediction. The proposed Denseformer can make use of Densenet and Transformer to learn more powerful features.

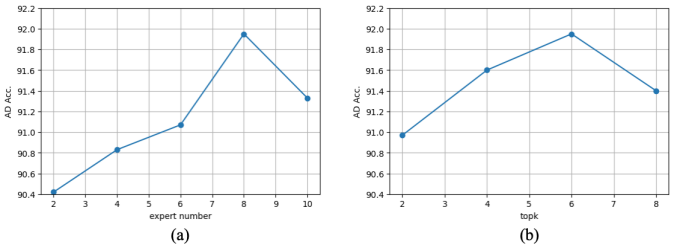
Second, we conduct experiments to evaluate the impact of DenseFormer blocks by increasing the number of DenseFormer blocks from 1 to 5. The results for AD diagnosis on the ADNI dataset are compared in TABLE III. We can see that increasing the number of DenseFormer blocks can improve model performance but increases computational cost and complexity. When the number of DenseFormer blocks is increased to 5, the performance improvement is marginal. Thus, to balance model capability and computational efficiency, our model includes four DenseFormer blocks.

Third, since the Feature Bridge facilitates the alignment and fusion of local and global features learned from the enseFormer, we conduct the experiments of AD diagnosis on the ADNI dataset by using DenseFormer backbone without MAE pre-training. The results with and without Feature Bridge are compared in TABLE IV. The DenseFormer without Feature Bridge is implemented by combining the features of Densenet and Transformer streams with a fully connected layer for final decision. The results show that the feature bridge can enhance the performance of AD diagnosis.

Effectiveness of MAE Self-Supervised Pre-training: In this experiment, we test the effectiveness of pre-training the DenseFormer (DenF) backbone using MAE self-supervised learning (denoted as "DenF+MAE") for AD diagnosis on ADNI dataset. We compare the Densenet, Vision Transformers and DenseFormer without and with MAE pre-training as well as the DenseFormer pre-training with different self-supervised learning methods: Rotation [37], MoCo [24] and DINO [26], which are denoted as "Densenet", "ViT", "Densenet+MAE", "ViT+MAE", "DenF+MAE", "DenF+Rota", "DenF+MoCo" and "DenF+DINO", respectively. These methods are implemented using their released source codes with our best efforts. The contrastive learning methods, DINO in [26] and MoCo in [24], employ various data augmentation techniques such as random cropping, color jittering, random rotations, and horizontal flipping. These augmentations are typically used

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT MODELS AND PRE-TRAINING METHODS FOR AD DIAGNOSIS ON ADNI DATASET

Method	Accuracy	Sensitivity	Specificity	AUC
Densenet	0.870	0.807	0.920	0.935
Densenet+MAE	0.873	0.820	0.919	0.934
ViT	0.691	0.621	0.745	0.721
ViT+MAE	0.820	0.692	0.920	0.900
DenF	0.887	0.833	0.930	0.937
DenF+Rota	0.890	0.871	0.905	0.937
DenF+MoCo	0.892	0.868	0.916	0.939
DenF+DINO	0.895	0.865	0.918	0.940
DenF+MAE	**0.907**	**0.884**	**0.925**	**0.946**



(a) (b)

Fig. 5. The effect of varying the number of experts (a) and the topk parameter (b) on AD classification accuracy using the ADNI dataset.

TABLE VI
PERFORMANCE COMPARISON OF SINGLE TASK LEARNING (STL), TASK WEIGHTED LOSS FUSION (TWL) [30], CURRICULUM LEARNING (CL) [38], GRAD-NORM (GN) [31], UNCERTAINTY WEIGHTED LOSSES (UWL) [32], AND MOE MULTI-TASK LEARNING (MOE) FOR AD, MCI, PD AND DE DIAGNOSIS AND AGE PREDICTION.

Task	Perf.	STL	CL	TWL	GN	UWL	**MoE**
AD vs. NC	ACC↑	0.909	0.879	0.851	0.882	0.911	**0.923**
	SEN↑	0.887	0.889	0.716	0.899	0.865	**0.932**
	SPEC↑	0.923	0.876	0.954	0.868	0.936	**0.915**
	AUC↑	0.948	0.951	0.803	0.950	0.951	**0.979**
MCI vs. NC	ACC↑	0.765	0.768	0.747	0.765	0.774	**0.789**
	SEN↑	0.690	0.680	0.668	0.735	0.749	**0.725**
	SPEC↑	0.824	0.827	0.811	0.625	0.757	**0.834**
	AUC↑	0.826	0.830	0.821	0.827	0.832	**0.836**
sMCI vs. pMCI	ACC↑	0.781	0.738	0.707	0.759	0.774	**0.818**
	SEN↑	0.651	0.612	0.516	0.659	0.645	**0.648**
	SPEC↑	0.911	0.852	0.882	0.852	0.892	**0.932**
	AUC↑	0.812	0.783	0.775	0.801	0.809	**0.819**
PD vs. NC	ACC↑	0.838	0.822	0.788	0.790	0.831	**0.851**
	SEN↑	0.930	0.881	0.865	0.837	0.953	**0.947**
	SPEC↑	0.745	0.718	0.708	0.741	0.701	**0.744**
	AUC↑	0.825	0.830	0.814	0.822	0.829	**0.843**
DE vs. NC	ACC↑	0.840	0.820	0.792	0.831	0.845	**0.861**
	SEN↑	0.776	0.812	0.868	0.694	0.748	**0.780**
	SPEC↑	0.891	0.835	0.721	0.920	0.917	**0.922**
	AUC↑	0.915	0.905	0.855	0.928	0.923	**0.936**
Age Prediction	MAE↓	2.660	2.762	2.851	2.781	2.730	**2.590**
	RMSE↓	3.378	3.519	3.611	3.515	3.425	**3.321**
	PCC↑	0.891	0.891	0.887	0.893	0.888	**0.899**

TABLE VII
COMPARISON OF MULTI-TASK METHODS FOR DOWNSTREAM TASKS ON ABIDE AND OASIS USING THE DENSEFORMER BACKBONE.

Dataset	Task	Perf.	STL	TWL	UWL	**MoE**
OASIS3	AD vs. NC	ACC↑	0.912	0.900	0.913	**0.928**
		SEN↑	0.865	0.857	0.861	**0.874**
		SPEC↑	0.948	0.942	0.950	**0.956**
		AUC↑	0.922	0.916	0.924	**0.930**
	FD vs. NC	ACC↑	0.866	0.830	0.887	**0.900**
		SEN↑	0.833	0.612	0.854	**0.750**
		SPEC↑	0.888	0.947	0.936	**0.980**
		AUC↑	0.889	0.878	0.915	**0.927**
ABIDE	ASD vs. NC	ACC↑	0.554	0.531	0.571	**0.583**
		SEN↑	0.904	0.885	0.906	**0.910**
		SPEC↑	0.238	0.206	0.245	**0.276**
		AUC↑	0.523	0.501	0.549	**0.557**

to generate positive and negative sample pairs for contrastive learning, specifically for natural images. In our implementation, these augmentations are adapted to the sMRI data and preserve critical diagnostic regions, such as random rotations within ±10 degrees and horizontal flips. In the Rotation method [37], the training images are randomly rotated by 0°, 90°, 180° and 270° along the axial plane, and the DenseFormer is trained with a cross-entropy loss to predict the rotation angle of input images. In the contrastive learning method MoCo [24], the DenseFormer is trained with a momentum contrastive loss to optimize the alignment of positive pairs and the divergence of negative samples for unsupervised feature learning. In the contrastive learning method DINO [26], the DenseFormer is trained within a teacher-student framework with a contrastive loss that aligns the student's output with the teacher's for augmented views of the same image, enhancing feature representation through self-supervised learning.

The results are compared in TABLE V. We can see that MAE pre-training slightly improves the performance of Densenet, while it significantly enhances ViT's performance. DenseFormer performs better than the Densenet and ViT. The self-supervised learning methods: Rotation, MoCo and DINO can improve the performance of the DenseFormer. Nevertheless, the DenseFormer with MAE has achieved the best performance with 2% improvement than that without MAE. The results demonstrate MAE self-supervised pre-training can improve performance of DenseFormer in learning features for downstream diagnosis tasks.

C. Effectiveness of MoE Multi-Task Learning

In this section, we conduct an ablation study to test the effectiveness of MoE Multi-Task Learning in our DenseFormer-MoE model. Experiments test the impacts of MoE parameters and Multi-Task Learning, and analyze computation costs and performance across various downstream datasets.

First, since the total and selected experts are key parameters for MoE multi-task learning, we test the impact of these parameters in our DenseFormer-MoE for AD diagnosis on the ADNI dataset. In the experiment, the number of experts is increased from 2 to 10, and the results are shown in Fig. 5 (a). We can see that the diagnosis accuracy is gradually improved to 91.95% by increasing the number of experts to 8, while it is slightly reduced to 91.33% with the number of experts increased to 10. These results indicate that adding more experts can enhance performance, but too many experts increase complexity and degrade performance. Thus, in the following experiments, the total number of experts is set to 8 for MoE multi-task learning. In addition, since the gating network of MoE selects the topk experts for the final decision, we change the number of selected experts and test the performance of AD diagnosis. The results are shown in Fig. 5 (b). We can see that the best performance is achieved with topk set to 6.

TABLE VIII
COMPARISON OF COMPUTATION COSTS IN BACKBONE PRE-TRAINING, MULTI-TASK FINE-TUNING AND TESTING.

Stage	Method	Training time (h)	Inference time (s)
Pre-training	DenseFormer+MAE	10.60	-
Fine-tuning	DenseFormer+TWL	2.07	0.1171
	DenseFormer+UWL	2.01	0.1172
	DenseFormer+MoE	**1.87**	**0.1279**

Second, we compare our MoE multi-task learning (MoE) with Single Task Learning (STL) and other multi-task learning methods including task weighted loss fusion (TWL) [30], Curriculum Learning(CL) [38], GradNorm (GN) [31], and Uncertainty Weighted Loss (UWL) [32]. The experiments are conducted on UKB, ADNI and PPMI datasets for brain age prediction and five classification tasks of AD vs. NC, MCI vs. NC, sMCI vs. pMCI, PD vs. NC and DE vs. NC, where DE merges the AD and PD into one degenerative disease. For a fair comparison, all experiments are conducted with the pre-trained DenseFormer backbone by MAE, followed by different task learning strategies. In STL, each task is fine-tuned separately with the same pre-trained backbone, limiting cross-task information sharing. TWL, which averages the task-specific losses, assumes equal importance across tasks, making it suitable when tasks are closely related. CL introduces tasks progressively, from easiest to hardest, to guide learning, which gradually refines the model. GN modulates gradient magnitudes to ensure balanced learning across tasks, addressing issues of task imbalance. UWL dynamically weights task losses based on uncertainty, enabling the model to focus on challenging tasks. TWL, CL, GN, and UWL try to improve multi-task learning by modifying the loss function. However, these loss-based methods may have suboptimal performance if tasks conflict or the gradients diverge. Our proposed method employs gating and expert networks to dynamically route task-specific information and adaptively select relevant features through the MoE mechanism, thereby effectively reducing conflicts between tasks. Results of five tasks are compared in TABLE VI. This adaptive feature selection enables our MoE to achieve the best performance with AUCs of 0.979, 0.836, 0.843, and 0.936 for AD, MCI, PD, and DE diagnosis, respectively. Additionally, we achieved the best AUC of 0.819 for sMCI vs. pMCI classification, and an MAE of 2.590 years for brain age prediction. The results show our MoE method can achieve superior performance across brain imaging tasks.

Third, we evaluate the computation costs of the proposed method, consisting of both the training and testing stages. The training stage includes DenseFormer backbone pre-training and MoE multi-task fine-tuning phases. With the pre-trained backbone, we compare the computation costs between the proposed MoE method and other multi-task methods: TWL, UWL. The computation costs of pre-training, fine-tuning and testing are compared in TABLE VIII. The experiments are conducted for the classification tasks of AD vs. NC, MCI vs. NC, sMCI vs. pMCI, PD vs. NC, and DE vs. NC on the NVIDIA 4090ti GPUs, with the batch size set to 6 for all methods to ensure a fair comparison of convergence time. The inference time is measured as the average inference time for ten input samples. From the results, we can see that the proposed MoE method is faster to converge than UWL and TWL for the fine-tuning stage, which is consistent with those in [13]. In the testing stage, although the proposed MoE method integrates multiple gating subnetworks and expert networks, its inference time is comparable to those of TWL and UWL, with a slight increase of 0.01s. The results show that MoE offers faster convergence in the fine-tuning without significantly affecting the inference speed.

TABLE IX
PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH OTHER METHODS FOR PREDICTION OF BRAIN AGE AND DIAGNOSIS OF AD AND PD BASED ON SMRI DATA FROM UKB, ADNI AND PPMI.

Dataset	Task	Method	ACC↑	SEN↑	SPEC↑	AUC↑
ADNI	AD vs. NC	Lian *et al.* [8]	0.903	0.824	0.965	0.951
		Liu *et al.* [2]	0.892	0.859	0.925	0.950
		Liu *et al.* [39]	0.876	0.846	0.900	0.922
		Chen *et al.* [9]	0.880	0.823	0.908	0.927
		Qiu *et al.* [3]	0.889	0.830	0.916	0.935
		Cai *et al.* [17]	0.855	0.724	0.891	0.905
		DenseFormer-MoE	**0.923**	**0.932**	**0.915**	**0.979**
PPMI	PD vs. NC	Camacho *et al.* [10]	0.835	0.914	0.672	0.778
		Qiu *et al.* [3]	0.752	0.819	0.726	0.795
		DenseFormer-MoE	**0.851**	**0.947**	**0.744**	**0.843**
			MAE↓	RMSE↓	PCC↑	
UKB	Age Prediction	He *et al.* [1]	3.09	3.91	0.852	
		He *et al.* [11]	2.67	3.36	0.893	
		Peng *et al.* [40]	2.64	3.34	0.896	
		Cai *et al.* [17]	2.73	3.72	0.871	
		DenseFormer-MoE	**2.59**	**3.32**	**0.899**	

Finally, to validate the effectiveness of our proposed method on a broader range of brain imaging datasets, we conduct multi-task classification experiments utilizing the pre-trained DenseFormer backbone on the ABIDE and OASIS3 datasets. These experiments encompass classifications of AD vs NC, Frontotemporal Dementia (FD) vs. NC, and Autism Dementia (ASD) vs. NC. The MoE method is compared against other single-task and multi-task methods: STL, TWL and UWL, as shown in TABLE VII. The results of ACC, SEN, SPE and AUC demonstrate that the MoE method consistently performs better than other methods across all tasks. Notably, the results of ASD are less satisfactory due to our method solely relying on the sMRI data from ABIDE, which lacks discriminatory power for Autism diagnosis.

D. Comparison With Other Methods

TABLE IX compares the proposed DenseFormer-MoE method with other recently published state-of-the-art approaches in the fields of AD diagnosis [2], [3], [8], [9], [17], [39], PD diagnosis [3], [10], and brain age prediction [1], [11], [17], [40] on UKB, ADNI and PPMI datasets. To ensure a fair comparison, we meticulously reproduce the methods proposed in [1], [3], [9], [11], [17], [39], [40] using their publicly available code on GitHub and employing the same training configurations on the same training and testing datasets. Although the method proposed in [10] does not provide publicly available source code, it employs a straightforward fully convolutional neural network model, which we reimplemented for comparison. For the methods [2], [8], which used training and testing datasets similar to ours, we directly use the results reported in their published papers for comparison.

These methods can be categorized based on their model structures into traditional machine learning-based methods [9], deep CNN-based methods [1]–[3], [8], [10], [39], [40], and hybrid deep learning-based methods [11], [17]. Among these, the hybrid architecture proposed in [11], which primarily uses CNNs as front-end feature extractors cascaded with Transformers, demonstrates superior performance over the CNN-based method [1] in the age prediction task. However, our proposed DenseFormer-MoE not only outperforms the traditional machine learning-based and CNN-based methods, but also surpasses hybrid structures [11], [17] that combine CNNs with Transformers. DenseFormer-MoE fully integrates the strengths of CNNs and Transformers, achieving a more effective feature extraction and representation for various tasks.

Additionally, the proposed DenseFormer-MoE achieves better performance across different tasks when compared with the single-task methods [1], [2], [8]–[11], [39], [40] and the multi-task methods [17] and [3]. Specifically, the multi-task method [17] evaluates the performance across multiple tasks by independently fine-tuning separate models for each task, rather than jointly optimizing them within a unified framework. Similarly, the multi-task method [3] focuses on multi-task learning within the same disease category (e.g., AD vs. NC and MCI vs. NC) using multimodal data. When we reproduced this method using only sMRI data and extended it to different diseases (e.g., AD and PD), the results indicated that its performance was less optimal for this broader application. In contrast, our DenseFormer-MoE achieves superior performance across all tasks compared to these single-task and multi-task methods. For AD diagnosis, our method achieves the highest ACC of 0.923 and an AUC of 0.979. For PD diagnosis, it achieves the best accuracy of 0.851 and a competitive AUC of 0.843. Furthermore, for brain age prediction, DenseFormer-MoE obtains the lowest MAE of 2.59, RMSE of 3.32, and the highest PCC of 0.899. These results demonstrate the advantages of DenseFormer-MoE in handling multiple downstream tasks through robust multi-task learning.

E. Discussion

In addition to disease diagnosis, interpretation of the DenseFormer-MoE model is also important for brain image analysis. First, to facilitate interpretation, we generate the grad attention maps on all patient subjects (i.e., AD or PD) from the testing set with the method [41]. Each value of the attention map indicates the importance of the corresponding voxel to disease diagnosis through gradient calculation. All the attention maps ignore the effect of negative gradients and are normalized to [0, 1]. To show the relevant regions to different diagnosis tasks, we further calculate the average attention map and overlay it on a template image (MNI_152_T1_1mm), as shown in Fig. 6. From Fig. 6 (a), we can see that the salient brain regions for AD diagnosis include the hippocampus, amygdala, parts of the temporal lobe, parahippocampal gyrus and ventricle etc., which are consistent with those of the previous studies [8], [42]. For PD diagnosis, the relevant regions, primarily the caudate nucleus and putamen, play the pivotal roles in motor control and cognitive functions, aligning

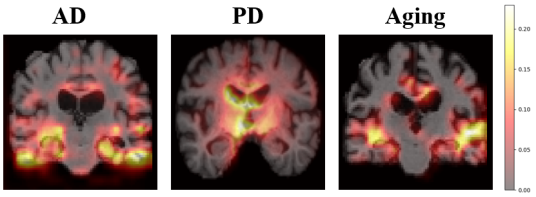


Fig. 6. The attention maps generated from the DenseFormer-MoE model using the GradCAM [41], with high values indicating more relevant regions associated with AD, PD and aging.

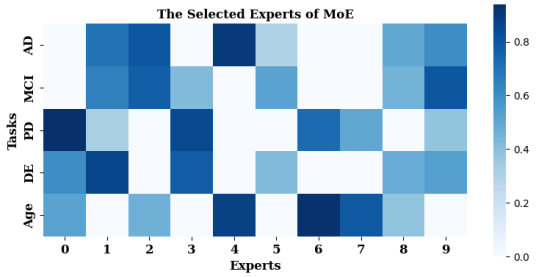


Fig. 7. The frequency map of 10 experts selected across different tasks in topk of the MoE module.

with the previous studies [43], [44]. For brain age prediction, the attention map indicates that aging is associated with the temporal lobe, thalamus, hippocampus, and brain atrophy, which are consistent with previous studies [11], [17].

Second, MoE-selected experts determine task results in multi-task learning. For interpretation, we analyze the selected experts by the MoE for different tasks. In our experiments, there are 10 experts in total, and the top 6 experts are selected by MoE for each test subject. The selection frequency of each expert for individual tasks is computed as shown in Fig. 7. Results show that different tasks have specialized experts, e.g., experts 1,2,4,5,8,9 are often selected for AD diagnosis. Since the tasks of diagnosing MCI and AD are closely related, their selected experts tend to overlap. Conversely, the tasks of diagnosing PD and AD are less correlated, resulting in the selection of distinct experts for each. The results indicate that various experts contribute differently to multiple tasks, and each task has its unique combination of experts.

V. Conclusion

This paper presents DenseFormer-MoE, a dense transformer foundation model integrating DenseNet, Transformer, MAE, and MoE to learn and fuse local and global sMRI features for multi-task brain image analysis. The DenseFormer backbone, combining DenseNet and vision Transformers, is pre-trained with MAE to enhance feature representation. MoE dynamically selects task-specific experts for multi-task learning, improving adaptability across different tasks. The results and comparison show the promising performances for various tasks including AD, MCI and PD diagnosis and brain age prediction. The proposed model only works on the sMRI data which limits its applications. In the future work, we will extend it to use multimodal data. Additionally, since the proposed model is data-driven, integrating prior knowledge could further enhance its interpretability and robustness.

VI. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62171283), the National key research and development program of China under Grant (2022YFC2503305 and 2022YFC2503302), the Fundamental Research Funds for the Central Universities (YG2023QNB27, YG2024LC11, YG2024QNA57, YG2024QNA58), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

[1] S. He, D. Pereira, J. D. Perez *et al.*, "Multi-channel attention-fusion neural network for brain age estimation: Accuracy, generality, and interpretation with 16,705 healthy mris across lifespan," *Medical Image Analysis*, vol. 72, p. 102091, 2021.

[2] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Medical image analysis*, vol. 43, pp. 157–168, 2018.

[3] S. Qiu, M. I. Miller, P. S. Joshi *et al.*, "Multimodal deep learning for alzheimer's disease dementia assessment," *Nature communications*, vol. 13, no. 1, p. 3404, 2022.

[4] Z. Huang, H. Lei, G. Chen *et al.*, "Parkinson's disease classification and clinical score regression via united embedding and sparse learning from longitudinal data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3357–3371, 2021.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[6] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[8] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 4, pp. 880–893, 2018.

[9] Z. Chen, Y. Liu, Y. Zhang, Q. Li, A. D. N. Initiative *et al.*, "Orthogonal latent space learning with feature weighting and graph learning for multimodal alzheimer's disease diagnosis," *Medical Image Analysis*, vol. 84, p. 102698, 2023.

[10] M. Camacho, M. Wilms, P. Mouches *et al.*, "Explainable classification of parkinson's disease using deep learning trained on a large multi-center database of t1-weighted mri datasets," *NeuroImage: Clinical*, vol. 38, p. 103405, 2023.

[11] S. He, P. E. Grant, and Y. Ou, "Global-local transformer for brain age estimation," *IEEE transactions on medical imaging*, vol. 41, no. 1, pp. 213–224, 2021.

[12] K. He, X. Chen, S. Xie *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF CVPR*, 2022, pp. 16 000–16 009.

[13] Z. Chen, Y. Shen, M. Ding *et al.*, "Mod-squad: Designing mixtures of experts as modular multi-task learners," in *Proc. IEEE/CVF CVPR*, 2023, pp. 11 828–11 837.

[14] P. M. Shah, A. Zeb, U. Shafi *et al.*, "Detection of parkinson disease in brain mri using convolutional neural network," in *Proc. Int. Conf. Autom. Comput. (ICAC)*. IEEE, 2018, pp. 1–6.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] H. Cai, Y. Gao, and M. Liu, "Graph transformer geometric learning of brain networks using multimodal mr images for brain age estimation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 456–466, 2022.

[18] Y. Xie, J. Zhang, C. Shen *et al.*, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Proc. MICCAI*. Springer, 2021, pp. 171–180.

[19] I. Higgins, L. Matthey, A. Pal *et al.*, "Beta-vae: Learning basic visual concepts with a constrained variational framework." *ICLR (Poster)*, vol. 3, 2017.

[20] L. Zhou, H. Liu, J. Bae *et al.*, "Self pre-training with masked autoencoders for medical image classification and segmentation," in *Proc. IEEE ISBI*, 2023, pp. 1–6.

[21] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in neural information processing systems*, vol. 33, pp. 12 546–12 558, 2020.

[22] S. Azizi, B. Mustafa, F. Ryan *et al.*, "Big self-supervised models advance medical image classification," in *Proc. IEEE/CVF ICCV*, 2021, pp. 3478–3488.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[24] K. He, H. Fan, Y. Wu *et al.*, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF CVPR*, 2020, pp. 9729–9738.

[25] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE/CVF CVPR*, 2021, pp. 15 750–15 758.

[26] M. Caron, H. Touvron, I. Misra *et al.*, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF ICCV*, 2021, pp. 9650–9660.

[27] P. Moeskops, J. M. Wolterink, B. H. M. Van Der Velden *et al.*, "Deep learning for multi-task medical image segmentation in multiple modalities," in *Proc. MICCAI*. Springer, 2016, pp. 478–486.

[28] R. Feng, Y. Cao, X. Liu *et al.*, "Chronet: A multi-task learning based approach for prediction of multiple chronic diseases," *Multimedia Tools and Applications*, pp. 1–15, 2022.

[29] S. Chelaramani, M. Gupta, V. Agarwal *et al.*, "Multi-task knowledge distillation for eye disease prediction," in *Proc. IEEE/CVF WACV*, 2021, pp. 3983–3993.

[30] A. Vlachostergiou, A. Tagaris, A. Stafylopatis *et al.*, "Multi-task learning for predicting parkinson's disease based on medical imaging information," in *Proc. IEEE ICIP*, 2018, pp. 2052–2056.

[31] Z. Chen, V. Badrinarayanan, C.-Y. Lee *et al.*, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. ICML*, 2018, pp. 794–803.

[32] A. Kendall, Y. Gal, R. Cipolla *et al.*, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE CVPR*, 2018, pp. 7482–7491.

[33] P. Zhu, Y. Sun, B. Cao *et al.*, "Task-customized mixture of adapters for general image fusion," in *Proc. IEEE/CVF CVPR*, 2024, pp. 7099–7108.

[34] Y. Zhang, R. Cai, T. Chen *et al.*, "Robust mixture-of-expert training for convolutional neural networks," in *Proc. IEEE/CVF ICCV*, 2023, pp. 90–101.

[35] F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter *et al.*, "Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank," *Neuroimage*, vol. 166, pp. 400–424, 2018.

[36] S. M. Smith, "Fast robust automated brain extraction," *Human brain mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[37] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[38] Y. Bengio, J. Louradour, R. Collobert *et al.*, "Curriculum learning," in *Proc. ICML*, 2009, pp. 41–48.

[39] S. Liu, A. V. Masurkar, H. Rusinek *et al.*, "Generalizable deep learning model for early alzheimer's disease detection from structural mris," *Scientific reports*, vol. 12, no. 1, p. 17106, 2022.

[40] H. Peng, W. Gong, C. F. Beckmann *et al.*, "Accurate brain age prediction with lightweight deep neural networks," *Medical image analysis*, vol. 68, p. 101871, 2021.

[41] R. R. Selvaraju, M. Cogswell, A. Das *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[42] C. Lian, M. Liu, Y. Pan, and D. Shen, "Attention-guided hybrid network for dementia diagnosis with structural mr images," *IEEE transactions on cybernetics*, vol. 52, no. 4, pp. 1992–2003, 2020.

[43] E. Adeli, G. Wu, B. Saghafi *et al.*, "Kernel-based joint feature selection and max-margin classification for early diagnosis of parkinson's disease," *Scientific reports*, vol. 7, no. 1, p. 41069, 2017.

[44] A. E. Taylor, J. A. Saint-Cyr, and A. E. Lang, "Frontal lobe dysfunction in parkinson's disease: The cortical focus of neostriatal outflow," *Brain*, vol. 109, no. 5, pp. 845–883, 1986.