MetaViT: Metabolism-Aware Vision Transformer for Differential Diagnosis of Parkinsonism with ^{18}F-FDG PET

Lin Zhao[1,2⋆], Hexin Dong[1,3], Ping Wu[4], Jiaying Lu[4], Le Lu[1], Jingren Zhou[1], Tianming Liu[2], Li Zhang[3], Ling Zhang[1], Yuxing Tang[1(✉)], Chuantao Zuo[4(✉)]

[1] DAMO Academy, Alibaba Group
[2] School of Computing, The University of Georgia, Athens, GA, USA
[3] Peking University, Beijing, China
[4] PET Center, Huashan Hospital, Fudan University, Shanghai, China

Abstract. Accurate and early differential diagnosis of parkinsonism (idiopathic Parkinson's disease, multiple system atrophy, and progressive supranuclear palsy) is crucial for informing prognosis and determining treatment strategies. Current automated differential diagnosis methods for ^{18}F-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) scans, such as convolutional neural networks (CNNs), often focus on local brain regions and do not explicitly model the complex metabolic interactions between distinct brain regions. These interactions, as reflected in FDG PET images, are keys for the differential diagnosis of parkinsonism. Vision transformer (ViT) models are promising in modeling such long-range dependencies, but they may overlook the local metabolic alternations and have not been widely adapted for 3D medical image classification due to data limitations. Therefore, we propose a novel metabolism-aware vision transformer (MetaViT), which uses self-attention and convolution to explicitly characterize both global and local metabolic interactions between interrelated brain regions. A masked image reconstruction task is introduced to guide the MetaViT model to focus on disease-related brain regions, addressing the scarcity of 3D medical imaging data and improving the trustworthiness and interpretability of the resulting model. The proposed framework is evaluated on a 3D FDG PET imaging dataset with 902 subjects, achieving a high accuracy of 97.7% in the differential diagnosis of parkinsonism and outperforming several state-of-the-art CNN and ViT-based approaches.

Keywords: PET · Parkinsonism · Early Differential Diagnosis · Transformer · Masked Image Reconstruction

1 Introduction

Idiopathic Parkinson's disease (IPD) is among the most common neurodegenerative disorders and has attracted the interest of both clinical and research trials

[⋆] This work was supported by Alibaba Group through Alibaba Research Intern Program. Email(✉): yuxing.t@alibaba-inc.com, zuochuantao@fudan.edu.cn for decades [1,7,9,19]. Accurate and early diagnosis of IPD plays a crucial role in determining potential therapeutic interventions and treatment outcomes [14]. However, it remains challenging due to the large overlap of IPD's symptoms with atypical parkinsonian syndromes like multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) [3,6,19], especially in the early stage. Recently, ^{18}F-fluorodeoxyglucose positron emission tomography (^{18}F-FDG PET) has demonstrated advantages in the differential diagnosis of parkinsonism prior to the development of brain structural damage, by revealing the brain glucose metabolism that indicates the abnormalities of the brain[23,19].

Based on ^{18}F-FDG PET, various computational tools have been developed to exploit the discriminative features from the metabolic patterns of the human brain for early and accurate differential diagnosis. For example, principal component analysis was employed to extract disease-specific features for machine learning algorithms, such as logistic regression [18] and scaled subprofile model [13,16]. Deep learning-based methods have also been widely adopted and applied in the differential diagnosis of parkinsonism [19,23]. For instance, a recent study proposed an IPD Diagnosis Network (PDD-Net) [19] based on a modified 3D deep residual convolutional neural network [5] to provide an end-to-end solution for automatic differentiation and achieved promising results than traditional machine learning-based methods.

Despite the wide adoption and success of the aforementioned techniques, current state-of-the-art differential diagnostic tools remain limited in the sense that they do not effectively model the complex metabolic interactions of interrelated brain regions, which are considered to crucially reflect the differences among parkinsonism in FDG PET images [19]. Most of the previous methods model the interactions in FDG PET images by adapting 3D convolutional neural networks (CNNs). However, CNNs do not explicitly characterize the global metabolic alternations of distinct brain regions, but gradually enlarge the receptive field from local to global through the integration of local information [15]. Recently, vision transformers (ViTs) have become increasingly popular and dominant in image recognition tasks, offering comparable or even superior performance as an alternative to CNNs [2,11]. ViTs divide the entire image into several smaller image patches and model their interactions to aggregate global information, and seem to compensate for the aforementioned shortcoming of CNNs. However, ViTs are inefficient in integrating the local information compared with CNNs such that the local metabolic alternations may not be well characterized and thus degenerate the performance. Meanwhile, due to the lack of inductive biases, optimization of ViTs requires much more training samples than CNNs [2,10], especially in 3D scenarios where the number of parameters of the model far exceeds the number of 3D samples. To overcome this limitation and take advantage of ViTs, a possible way is to rely on fine-grained annotations and to perform the segmentation simultaneously with additional pixel/voxel-level supervision [22]. Nonetheless, PET images measure the metabolic activity of the human brain, whereas segmentation tasks are usually performed on anatomical structures. Therefore, there remains a desperate need for a general and effective framework tailored for the differential diagnosis of parkinsonism using FDG PET images, which combines the advantages of CNNs and ViTs and overcomes their respective drawbacks.

Motivated by this, we propose a novel metabolism-aware vision transformer (MetaViT), as shown in Fig. 1, for accurate differential diagnosis of parkinsonism. Our MetaViT model is specifically designed to explicitly describe both global metabolic interactions of interrelated brain regions and local metabolic alternations within small specific brain regions. To this end, we employ convolution operations to integrate local spatial information, and inter-patch voxel-wise self-attention operations as well as feed-forward layers to mimic the global interactions of metabolism in the brain. Moreover, prior knowledge from the nuclear radiologist is integrated during the model training to guide the MetaViT model to focus more on the brain's regions of interest (ROIs) that are highly correlated with potential metabolic abnormalities through a masked image reconstruction task (Fig. 1(b)). This unique design provides additional supervision for model optimization and compensates for the limited amount of data. In addition, a self-learning strategy is implemented to take advantage of additional data with noisy (clinically possible diagnosis) labels. Experiments on a 3D FDG PET imaging dataset (n=902) demonstrate the validity and effectiveness of the proposed framework, as well as its superior performance over state-of-the-art CNN and ViT-based methods. We also find that the integration of clinical prior knowledge improves the trustworthiness and interpretability of the resulting model.

The main contributions of our work are summarized as follows:

- We propose a novel metabolism-aware vision transformer (MetaViT) that explicitly characterizes both global and local interactions of glucose metabolism in the brain, demonstrating superior performance than CNN- and ViT-based baselines for 3D FDG PET image classification.
- We propose a masked image reconstruction task to guide MetaViT to focus more on disease-related brain regions that reflect the metabolic changes, which not only compensates for the lacking of sufficient 3D training images but also improves the model's trustworthiness and interpretability.
- Our framework provides a feasible solution to take advantage of ViT for 3D medical image classification, achieving state-of-the-art performance on the differential diagnosis of parkinsonism in FDG PET imaging, suggesting great promise in integrating prior knowledge for 3D medical image classification.

2 Methods

2.1 Metabolism-Aware Vision Transformer

The complex metabolic interactions of interrelated brain regions in the human brain are suggested to be different among parkinsonism in FDG PET images [19]. However, it has not been effectively characterized in previous studies due to methodological limitations such as the incompetence of CNNs in modeling long-range and global dependencies explicitly. To address this problem, in this subsection, we propose a novel Metabolism-Aware Vision Transformer (MetaViT

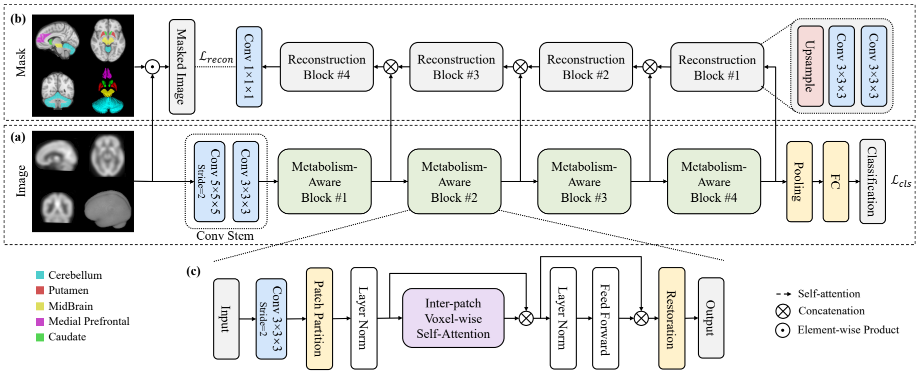

Fig. 1. Illustration of the proposed framework. (a) The architecture of the MetaViT with one convolution stem, four consecutive metabolism-aware blocks, one pooling and one fully-connected layer. (b) Masked image reconstruction branch consisting of four reconstruction blocks. Each reconstruction block is composed of two convolutional layers and one upsampling layer. (c) The constitution of the metabolism-aware block. The convolution integrates the local metabolic alternations while inter-patch voxel-wise self-attention interacts with those global alternations in a data-driven manner.

tailored for this objective. As illustrated in Fig. 1(a), our design of MetaViT follows a typical hierarchical scheme of CNNs (e.g., ResNet [5]) which consists of a convolutional stem and four consecutive metabolism-aware blocks (Fig. 1(c)) to model the metabolic interactions of interrelated brain regions.

Formally, given an input tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times D\times C}$ where H, W, D, C represent the height, width, depth, and the number of channels, respectively, we firstly model the local dependencies of metabolic alternations by applying a $3 \times 3 \times 3$ convolution with a stride of 2 (Fig. 1c, blue rectangular). We visualize this process in the left panel of Fig. 2. Each voxel (denoted by red cube) in the resulting tensor $\mathbf{X}_{conv} \in \mathbb{R}^{H/2\times W/2\times D/2\times C'}$ integrates the metabolic information of its neighbor voxels (denoted by green cube). Then, the long-range and global dependencies are characterized to enable the interactions across the whole brain. To do so, we introduce an inter-patch voxel-wise self-attention to model long-range dependencies. Specifically, $\mathbf{X}_{conv}$ is divided into N non-overlapping 3D patches $\mathbf{X}_{conv} \in \mathbb{R}^{V\times N\times C'}$ (Fig. 2, middle panel) where $V = h \times w \times d$ is the number of voxels within a patch, h,w,d are the height, width, depth of a patch, respectively, and $N = \frac{H\times W\times D}{8V}$ is the number of patches. For each voxel position $v \in \{1, \cdots, V\}$ within a patch, we then apply multi-head self-attention f_{MSA} and multilayer perceptron f_{MLP} to obtain the $\mathbf{X}_{trans} \in \mathbb{R}^{V\times N\times C'}$:

$$\mathbf{X}'_{trans}(v) = f_{\mathrm{MSA}}(\mathrm{LN}(\mathbf{X}_{conv}(v))) + \mathbf{X}_{conv}(v) \tag{1}$$

$$\mathbf{X}_{trans}(v) = f_{\mathrm{MLP}}(\mathrm{LN}(\mathbf{X}'_{trans}(v))) + \mathbf{X}'_{trans}(v) \tag{2}$$

where $\mathrm{LN}(\cdot)$ represents the layer normalization. $\mathbf{X}_{trans}$ can be restored as $\mathbf{X}_H \in \mathbb{R}^{H/2\times W/2\times D/2\times C'}$, which is the output for metabolism-aware block. Notably,

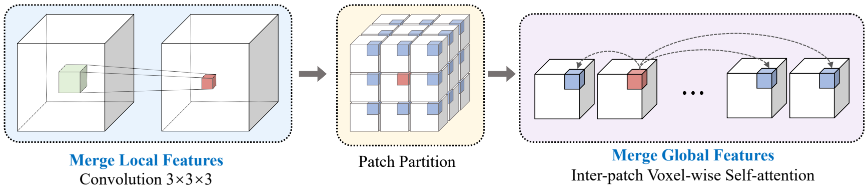

Fig. 2. Illustration of the proposed metabolism-aware design which integrates both local and global information. **Left**: local features are merged through convolution. The resulting voxel (red cube) integrates information from its local neighbors (green cube). **Middle**: the output of the convolution is divided into non-overlapping patches. **Right**: the final resulting voxel (red cube) integrates global information using inter-patch voxel-wise self-attention from voxels of the same position from other patches (blue cubes), which already contain their local information.

$\mathbf{X}_{conv}(v)$ integrates the local metabolic information through the 3D convolution operation, and $\mathbf{X}_{trans}(v)$ encodes the global metabolic information across N patches for the v^{th} position in a patch. As illustrated in the right panel of Fig. 2, the voxel in a patch (denoted by the red cube) integrates the information from the same position voxel of other patches (denoted by blue cubes) which already merge their local information. In this way, each voxel in $\mathbf{X}_{trans}$ can interact and infuse the metabolic information from all voxels of FDG PET images in a data-driven manner, which is congruent with our objective. Compared with previous methods, this approach explicitly and efficiently models the metabolic interactions across the brain rather than being limited to local areas. The interrelations of different brain regions are also delineated implicitly through this process.

The MetaViT can be optimized by minimizing the loss function $\mathcal{L}_{cls}$, which is the cross-entropy between the ground truth label y and the predictions $\hat{y}$:

$$\mathcal{L}_{cls} = -\sum_{i} y_i log(\hat{y_i}) \tag{3}$$

2.2 Masked Image Reconstruction

Optimization of the vision transformer usually requires a large number of training samples. However, in 3D medical imaging scenarios, the amount of data is usually scarce and limited. The intuition is to perform the segmentation task simultaneously, as it would additionally include pixel-level supervision. However, PET images reflect the metabolism rather than the anatomy of the human brain, and hence, the segmentation task is not feasible for our objective.

In this subsection, we introduce a novel masked image reconstruction task to integrate the clinical prior knowledge to guide the model optimization. The main idea of this approach is to utilize the intermediate features from each block of MetaViT to reconstruct the masked original image $X_M = X \odot M$, where M is a binary mask indicating the ROIs of disease-related brain regions and $\odot$ represents the element-wise product. In this way, the model is enforced to learn more features from the brain's ROIs that are highly related to potential metabolic abnormality for a better reconstruction. It actually accelerates the model optimization by utilizing prior knowledge instead of learning from a large amount of data by the model itself.

Specifically, as illustrated in Fig. 1(b), the masked image reconstruction branch consists of several reconstruction blocks. Each block is composed of two $3 \times 3 \times 3$ convolutions and an up-sampling operation with a scale factor of 2. The output features from each metabolic-aware block are firstly concatenated with the output from the previous reconstruction block (except the last one) and then fed into the next reconstruction block. The final output is processed by a $1 \times 1 \times 1$ convolution and considered as the reconstruction of the masked image. The optimization of the masked image reconstruction branch can be performed by minimizing the mean squared error (MSE) between the original masked image X_M and the reconstructed one X'_M:

$$\mathcal{L}_{recon} = \|X_M - X'_M\|_F^2 \tag{4}$$

The final loss function $\mathcal{L}$ for optimizing the whole framework is:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{recon} \tag{5}$$

where λ is a scaling factor to balance the two terms.

2.3 Self-learning for Noisy-labeled Data

In clinical practice, in addition to patient cohorts with definite diagnoses and confirmative diagnoses with follow-up, there are additional Parkinsonian patients with possible diagnoses, where physicians are unable to reach a clinically assured/definite differential diagnosis (considered as "noisy"-labeled data), especially for those in very early stages. To exploit the noisy-labeled data, we employ the self-learning strategy [20,8]. Different from previous self-learning methods of having one teacher, we introduce two additional teachers to benefit from different models' expertise and from the knowledge of the possible diagnoses.

Specifically, we first train the MetaViT model based on masked image reconstruction and a 3D ResNet-18 model [5] on the data with clinically definite diagnosis (ground truth-labeled data). After training, these two teachers are applied to noisy-labeled data to generate predicted labels, and the final pseudo label for these data is obtained by voting among two predicted labels and the noisy label itself. As such, the student model is subsequently forced to learn from the additional knowledge provided by the ResNet-18 and clinically possible diagnoses. Moreover, the student's knowledge is expanded through learning more data that contains more parkinsonism variations, allowing the student to learn beyond his teachers to be capable of classifying more challenging cases.

Lastly, both the noisy-labeled data with the final pseudo labels and the original training data with ground truth labels are used to train the final student model from scratch. Note that such a training strategy is found to be more effective than initializing the student model with the teacher model or first pretraining on noisy-labeled data and then finetuning on original training data [20].

3 Experiments and Results

3.1 Dataset and Pre-Processing

In this study, we adopt the Huashan Parkinsonian PET Imaging (HPPI) dataset (approved by its Institutional Review Board) with 528 subjects (IPD: 277, MSA: 149, PSP: 102; with clinically definite diagnoses) for evaluating the proposed framework and another 374 subjects with clinically possible diagnoses (noisy labels) for self-learning. The brain emission data were acquired 60 minutes after the injection of approximately 185 MBq ^{18}F-FDG and lasted for 10 minutes. After the attenuation correction with low-dose CT before the emission scan, scatter, dead time, and random coincidences, the ^{18}F-FDG PET data were reconstructed by the ordered subset expectation maximization method. The PET images were then registered to standard MNI 152 space and smoothed by a 3D Gaussian filter (σ=5 mm). Besides, we cropped the PET image with a size of $80 \times 96 \times 80$ and normalized it with a mean of 0 and a standard deviation of 1. We invited an expert nuclear radiologist with >10 years of experience in diagnosing parkinsonism to manually annotate the ROIs on a T1w template in $1 \times 1 \times 1mm$ resolution in MNI 152 space (Fig. 1(b)). The ROIs were then registered to the preprocessed individual PET images and smoothed by a 3D Gaussian filter (σ=5mm) as the final mask for masked image reconstruction.

3.2 Implementation Details and Compared Methods

In our experiments, the proposed model and all compared baselines are trained for 50 epochs with a batch size of 16. We use the AdamW optimizer [12] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a cosine annealing learning rate scheduler with initial learning 10^{-4} and 5 warm-up epochs. The scale factor λ in Eq. (5) is set as 1. The framework is implemented with PyTorch (`https://pytorch.org/`) deep learning library and the model is trained on a single NVIDIA V100 GPU. We conduct nested five-fold cross-validation. For self-learning, the predicted labels of the noisy-labeled data utilized in each fold are predicted by the MetaViT and 3D ResNet-18 models of the same fold. Note that the pseudo-labeled data is only used in the model training, not in validation.

The compared baselines can be roughly categorized as CNN-based and ViT-based methods. CNN-based category contains PDD-Net [19], two ResNets (ResNet-18 and 50) [5], and visual attention network (VAN) [4]. It is noted that VAN consists of purely convolution operations while the overall architecture follows the design of ViT, which achieves state-of-the-art performance in natural image recognition tasks. The ViT-based class contains a vanilla ViT-Tiny model, Swin Transformer model [11], and two other methods designed especially for small datasets by including the inductive biases from CNN: T2T-ViT [21]and ViT-SD [10]. Considering the limited amount of data and the overfitting problem, we only report the tiny- or small-scale setting with fewer parameters for each model and re-implement them to fit the 3D inputs. The permutation test with 1,000 permutations is used to perform statistical significance ($p<0.05$ indicating significance) comparisons of classification accuracies between different methods.

3.3 Differential Diagnosis Results

In this subsection, we report the performance of the proposed framework in differential diagnosis to demonstrate its effectiveness and superiority. In Table 1, we report the average accuracy with standard deviation, F1 score, sensitivity, specificity, and the number of trainable parameters for each compared baseline and the proposed method. It is observed that CNN-based methods outperform ViT-based methods, among which vanilla ViT has the worst performance (81.3%). This is probably because inductive biases in CNNs are important for scenarios with limited samples. Because the vanilla ViT model requires much more training samples, in our task, it leaves a large performance margin. The performance of Swin-T is also inferior to CNN-based methods. We assume that the small window partition of the Swin Transformer may be inefficient in modeling the metabolic interactions of the human brain. T2T-ViT and ViT-SD explicitly introduce inductive biases in their architecture design, obtaining great improvement over vanilla ViT (around 10%). Our proposed framework achieves state-of-the-art performance in terms of all evaluation metrics compared to baselines. The accuracy of our framework is significantly (p<0.001) higher than all compared methods. Notably, the sensitivity of PSP is improved from around 85% to 96%. In clinical practice, PSP is relatively rare and can be easily misdiagnosed as IPD, while PSP is more malignant than IPD and has totally different treatment protocols. The proposed framework can better support radiologists in making an accurate diagnosis for rare but severe diseases such as PSP. Overall, these results demonstrate the superiority of the proposed framework compared to baselines.

Table 1. The performances of the proposed method and compared baselines for differential diagnosis of parkinsonism. The average accuracy (Acc.) with standard deviation (std), F1 score, sensitivity and specificity over nested five-fold cross-validation are reported. Model sizes in terms of the number of trainable parameters (in M) are shown in the last column. ∗ indicates p-value<0.001 compared to the reference (i.e., **Ours**).

Methods	Acc.±std (%)	F1 Score (%)			Sensitivity (%)			Specificity (%)			Params
		IPD	MSA	PSP	IPD	MSA	PSP	IPD	MSA	PSP	(M)
PDD-Net [19]	92.8±2.8*	94.4	95.0	84.3	95.6	94.8	82.7	92.5	98.2	97.2	6.0
ResNet-18 [5]	93.7±1.4*	95.1	95.7	86.1	95.5	95.8	84.1	93.9	98.1	97.4	33.2
ResNet-50 [5]	92.6±1.3*	93.8	95.5	84.4	94.2	94.0	85.8	92.8	99.0	96.0	46.2
VAN-T [4]	92.6±2.1*	94.5	94.8	83.0	96.8	94.0	78.9	91.3	98.1	97.7	5.6
ViT-T [2]	81.3±4.6*	84.7	85.3	63.5	89.5	84.3	56.8	76.7	95.0	95.3	6.2
Swin-T [11]	86.4±4.0*	88.1	88.3	74.7	89.6	88.8	72.5	86.1	97.1	93.8	29.4
T2T-ViT-12 [21]	91.5±3.7*	92.5	94.1	84.8	93.4	93.1	84.6	91.3	98.1	96.5	13.0
ViT-SD-T [10]	92.2±2.3*	93.9	94.8	83.6	93.4	95.3	84.5	94.0	97.6	96.0	17.3
Ours-Backbone	95.5±1.8*	95.8	97.2	90.6	97.4	96.0	88.3	93.6	**99.5**	98.6	8.9
Ours	**97.7±1.0**	**97.9**	**98.3**	**96.1**	**98.1**	**97.9**	**96.1**	**97.7**	**99.5**	**99.1**	18.3

Table 2. Ablation study. The average accuracy with standard deviation (std), p-value, F1 score, sensitivity and specificity over nested five-fold cross-validation with different configurations. Abbreviations: B: backbone; EF: early fusion; IR: image reconstruction; MIR: masked image reconstruction; SL: self-learning; SLE: self-learning with ensemble.

Methods	Acc.±std (%)	p	F1 Score (%)			Sensitivity (%)			Specificity (%)		
			IPD	MSA	PSP	IPD	MSA	PSP	IPD	MSA	PSP
a) B	95.5±1.8	Ref.	95.8	97.2	90.6	97.4	96.0	88.3	93.6	**99.5**	98.6
b) +EF	95.5±1.1	0.99	95.9	97.5	90.2	97.8	96.5	86.7	93.1	**99.5**	98.8
c) +IR	95.3±2.6	0.97	95.8	96.2	91.3	97.8	94.9	88.7	93.4	99.2	98.8
d) +MIR	96.4±1.6	0.06	96.7	97.7	93.2	97.8	96.8	91.8	95.3	**99.5**	98.8
e) +MIR+SL	97.0±1.5	0.05	97.4	98.0	94.2	97.8	**97.9**	93.6	96.9	99.2	98.8
f) +MIR+SLE	**97.7±1.0**	<0.01	**97.9**	**98.3**	**96.1**	**98.1**	**97.9**	**96.1**	**97.7**	**99.5**	**99.1**

3.4 Ablation Study

In this subsection, we conduct ablation experiments to validate the efficacy of each component in the proposed framework. We report the averaged accuracy, F1 score, sensitivity, and specificity over nested five-fold cross-validation for different configurations in Table 2. Compared with the baseline methods in Table 1, our a) MetaViT backbone outperforms both CNN-based and ViT-based methods, suggesting the advantages of explicitly modeling the metabolic interactions of interrelated brain regions. To evaluate the effectiveness of the masked image reconstruction task, we compare it with two approaches: b) early fusion fuses the mask as an additional channel for the input image, which is the simplest way to integrate prior knowledge with the mask; c) image reconstruction reconstructs the original image rather than the masked image. It is observed that the early fusion strategy is of no help to the diagnosis performance and the original image reconstruction even degenerates the accuracy. In contrast, the d) masked image reconstruction task significantly (p=0.06) improves the accuracy from 95.5% to 96.4%, implying the effectiveness of integrating prior knowledge into the model training. We also observe that self-learning with an additional 374 subjects further contributes to the diagnosis accuracy: f) our final configuration, which generates the pseudo label based on the ensemble of three teachers (MetaViT, ResNet-18, and clinically possible diagnosis), is better than e) the self-learning process only with one teacher (MetaViT). This indicates that self-learning with label ensemble, i.e., learning from multiple teachers, could be a useful strategy to leverage the noisy-labeled data.

3.5 Interpretation of Model Reasoning

We generate the saliency maps of the proposed framework (i.e., MetaViT based on masked image reconstruction and self-learning) using the full-gradient method [17] for interpreting the model reasoning. It assigns a correlation score to each pixel of the original image to show the numerical contribution to the model's

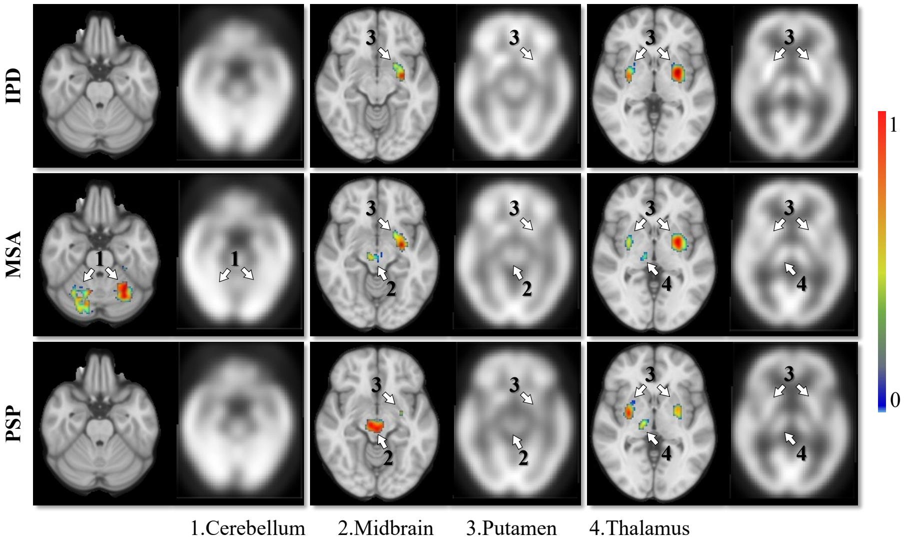

Fig. 3. Visualization of average saliency maps of patients with idiopathic Parkinson's disease (IPD), multiple system atrophy (MSA), and progressive supranuclear palsy (PSP) in the testing sets. Each row shows three axial slices of registered MRI and FDG PET averaged using all the subjects with the corresponding disease. We highlight the characteristic regions contributing to the above diseases based on our model. The color corresponds to the correlation score indicating the contribution of a region for each disease (red regions correspond to high scores). The arrows pointed to the most salient brain regions, including 1: cerebellum, 2: midbrain, 3: putamen, 4: thalamus.

decision-making. In Fig. 3, we illustrate average saliency maps (fused with T1-w modality template) and average FDG PET images of patients with IPD, MSA, or PSP in the testing cohort. Our model identifies four brain regions that are associated with the above diseases: the cerebellum, midbrain, putamen, and thalamus. In particular, our model focuses on the putamen for the prediction of all three diseases. Thalamus is linked to both MSA and PSP, while the cerebellum and midbrain are mainly associated with MSA or PSP, respectively. All these brain regions are either correlated with IPD/MSA/PSP pathology or confirmed by previous studies to be related to metabolic changes in IPD/MSA/PSP [18,19].

3.6 Limitations

One limitation of this study could be that it focuses on the major parkinsonian syndromes [18,19]. The ability to detect rare types of disorders (e.g., dementia with Lewy bodies, corticobasal degeneration) might be important in real-world applications. Additionally, although the 3D FDG PET imaging dataset used in this study is one of the largest patient cohorts in the literature, the sample size may still be small for deep learning analysis. More training data may further improve our model's performance. Finally, this study only includes patients from a single center, and it is unclear whether the results can be generalized to other centers with different patient populations.

4 Conclusion

This work presents a novel computational framework for accurate differential diagnosis of parkinsonian syndromes. Our MetaViT design emphasizes modeling the metabolic interactions of interrelated regions in the human brain, demonstrating superior performance than the compared CNN-based and ViT-based baselines. The proposed masked image reconstruction task leverages the clinical prior knowledge of disease-related brain regions to guide the model training, compensating for the lacking of sufficient 3D training samples with improved performances. Exploiting noisy-labeled data based on self-learning with label ensemble further improves the diagnosis accuracy. Our framework not only contributes to the accurate differential diagnosis of parkinsonism but also provides a feasible solution to take advantage of powerful ViT for 3D medical image classification, which has not yet been extensively studied due to limited 3D data. Our study would inspire future work to transpire on integrating prior knowledge into the deep model design and training to improve the reliability and transparency of medical imaging applications.

References

1. Braak, H., Rüb, U., Gai, W., Del Tredici, K.: Idiopathic parkinson's disease: possible routes by which vulnerable neuronal types may be subject to neuroinvasion by an unknown pathogen. Journal of neural transmission **110**(5), 517–536 (2003)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Eckert, T., Sailer, M., Kaufmann, J., Schrader, C., Peschel, T., Bodammer, N., Heinze, H.J., Schoenfeld, M.A.: Differentiation of idiopathic parkinson's disease, multiple system atrophy, progressive supranuclear palsy, and healthy controls using magnetization transfer imaging. Neuroimage **21**(1), 229–235 (2004)
4. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. arXiv preprint arXiv:2202.09741 (2022)
5. 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. pp. 770–778 (2016)
6. Hughes, A.J., Daniel, S.E., Ben-Shlomo, Y., Lees, A.J.: The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. Brain **125**(4), 861–870 (2002)
7. Hughes, A.J., Daniel, S.E., Kilford, L., Lees, A.J.: Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases. Journal of neurology, neurosurgery & psychiatry **55**(3), 181–184 (1992)

8. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with AlphaFold. Nature **596**(7873), 583–589 (2021)
9. Kish, S.J., Shannak, K., Hornykiewicz, O.: Uneven pattern of dopamine loss in the striatum of patients with idiopathic parkinson's disease. New England Journal of Medicine **318**(14), 876–880 (1988)
10. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
13. Matthews, D.C., Lerman, H., Lukic, A., Andrews, R.D., Mirelman, A., Wernick, M.N., Giladi, N., Strother, S.C., Evans, K.C., Cedarbaum, J.M., et al.: Fdg pet parkinson's disease-related pattern as a biomarker for clinical trials in early stage disease. NeuroImage: Clinical **20**, 572–579 (2018)
14. Pagan, F.L.: Improving outcomes through early diagnosis of parkinson's disease. American Journal of Managed Care **18**(7), S176 (2012)
15. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems **34**, 12116–12128 (2021)
16. Spetsieris, P.G., Ma, Y., Dhawan, V., Eidelberg, D.: Differential diagnosis of parkinsonian syndromes using pca-based functional imaging features. Neuroimage **45**(4), 1241–1252 (2009)
17. Srinivas, S., Fleuret, F.: Full-gradient representation for neural network visualization. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
18. Tang, C.C., Poston, K.L., Eckert, T., Feigin, A., Frucht, S., Gudesblatt, M., Dhawan, V., Lesser, M., Vonsattel, J.P., Fahn, S., et al.: Differential diagnosis of parkinsonism: a metabolic imaging study using pattern analysis. The Lancet Neurology **9**(2), 149–158 (2010)
19. Wu, P., Zhao, Y., Wu, J., Brendel, M., Lu, J., Ge, J., Bernhardt, A., Li, L., Alberts, I., Katzdobler, S., et al.: Differential diagnosis of parkinsonism based on deep metabolic imaging indices. Journal of Nuclear Medicine (2022)
20. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020)
21. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 558–567 (2021)
22. Zhang, L., Wen, Y.: A transformer-based framework for automatic covid19 diagnosis in chest cts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 513–518 (2021)
23. Zhao, Y., Wu, P., Wang, J., Li, H., Navab, N., Yakushev, I., Weber, W., Schwaiger, M., Huang, S.C., Cumming, P., et al.: A 3d deep residual convolutional neural network for differential diagnosis of parkinsonian syndromes on 18 f-fdg pet images. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 3531–3534 (2019)