AD-NeXt: Efficient Alzheimer's Disease Classification with Self-Supervised Pretraining

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Processing volumetric MRI data using conventional 3D convolutional networks demands substantial computational resources. To alleviate computational burden, this work introduces AD-NeXt, an adaptation of ConvNeXt-V2 architecture for MRI processing and Alzheimer's disease (AD) classification. We propose four architectural variants of different sizes and evaluate them on Alzheimer's disease classification datasets. Furthermore, to mitigate the limited availability of labeled medical imaging data for AD and enhance model performance, we employ self-supervised pre-training with masked autoencoders and sparse convolutions on unlabeled data from healthy individuals. Our models are compared against state-of-the-art architectures, with the base AD-NeXt model achieving superior classification performance while significantly reducing multiply-accumulate operations (MACs). Our code and pretrained model checkpoints are available at `https://anonymous.4open.science/r/convnext-for-ad-paper-Paper-2044`

Keywords: ConvNeXt · Alzheimer's Disease · Masked Autoencoders

1 Introduction

Deep learning has brought an automated disease diagnosis with unprecedented accuracy in medical imaging. However, it heavily depends on large-scale labeled datasets, which are scarce in the medical domain due to high annotation costs and the rarity of certain diseases. This data scarcity is particularly prominent in Alzheimer's disease (AD), where publicly available datasets of structural MRI are limited, hindering the development of robust vision models. One way to mitigate this challenge is to leverage larger unlabeled data through self-supervised learning to reduce the dependency on the annotations.

This approach was explored by Kunanbayev *et al.* [11] by pre-training Vision Transformers [4] on the MRI images from different clinical domains using Masked Autoencoders [6], and then fine-tuning on the smaller datasets with AD images. This technique yielded promising results, improving ViT's prediction accuracy by 12-14%. Despite advantages such as long-range dependencies [19] and lower floating-point operations per second (FLOPs) in processing 3D images compared to convolutional neural networks (CNNs) [21], transformers remain highly data-dependent. Consequently, they fail to match CNN performance in AD classification. [5].

Motivated by the strengths and limitations of both architectures, ConvNeXt [13] models were proposed as an enhancement over conventional CNNs. It utilizes depthwise convolutions [2] as building blocks, which, through weighted sum operations, resemble ViT's self-attention mechanism while being computationally more efficient than standard CNNs. In the follow-up work by Woo *et al.* [20] a Global Response Normalization (GRN) layer was added to ConvNeXt, which enables diverse feature representation and enhances the model's capacity through Masked Autoencoder pre-training with sparse convolutions [3]. Potentially, medical imaging practitioners could leverage ConvNeXt-inspired models, which can be computationally more efficient and achieve better performance when pre-trained on sufficient unlabeled data.

ConvNeXt models have been adapted for 3D image segmentation and classification, as seen in 3D-UX-Net [12] and MedNeXt [15]. However, compared to their 2D counterparts, these models are computationally expensive and relatively shallow due to their limited number of layers.

To address these challenges, we propose AD-NeXt, a variant of the ConvNeXt-V2 architecture [20] tailored for Alzheimer's disease (AD) classification. It achieves the depth of 3D models while maintaining significantly higher efficiency than traditional 3D CNNs. Additionally, we utilize self-supervised learning through masked autoencoder pre-training to enhance the performance of the model. Thus, we can list the contributions of the paper as follows:

- We replace 2D depthwise convolutions in ConvNeXt-V2 with their 3D counterparts. After optimal architecture search, we implement plain average pooling instead of strided convolution for downsampling. Additionally, we boost model performance using sparse convolutions and Masked Autoencoder pre-training on out-of-distribution (OOD) datasets of healthy individuals.
- We introduce multiple variations of our model from tiny to large, and release their pre-trained weights for the medical imaging community.
- Our pre-trained base model surpasses existing models in AD classification while using 29 times fewer multiply-accumulate operations[1] (MACs) than the second-best ResNet101 model.

2 Method

2.1 Datasets

In this work, experiments are conducted on publicly available datasets. For pre-training, IXI[2], HCP [18] containing MRI scans of only healthy subjects were utilized. In contrast, AD classification experiments were conducted on baseline collections of ADNI1 and ADNI2 of the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[3][10]. All MRI scans were pre-processed using HD-BET

[1] 1 GFLOPs = 2 GMACs [16]

[2] `https://brain-development.org/ixi-dataset`

[3] `https://adni.loni.usc.edu`

[9] for brain extraction. Additionally, we applied N4 bias field correction on HCP images using N4ITK [17] before brain extraction. More information about the datasets can be found in Table 1.

Table 1: Dataset details. $P_m(\%)$ indicates the share of the majority class. CN, Cognitively normal; AD, Alzheimer's disease

Dataset	**Magnet strength**	**AD**	**CN**	P_m **(%)**
IXI	1.5T/3T	–	581	–
HCP	3T	–	1113	–
ADNI1	1.5T	192	229	54.4
ADNI2	3T	159	201	55.8

2.2 Experimental setup

In all experiments, MRI scans are resampled to a voxel spacing of $1.75 \times 1.75 \times 1.75$, followed by foreground cropping, intensity normalization, and resizing to $128 \times 128 \times 128$ using the MONAI framework [1]. However, unlike ConvNeXt-V2 [20], during the pre-training stage, only geometric transformations such as random flipping and rotation are applied due to the smaller dataset size. In contrast, during fine-tuning, additional transformations such as random intensity scaling, shifting, and Gaussian noise were applied.

The experimental procedure is as follows. First, the base ConvNeXt-V2 model, adapted for 3D input, was systematically ablated across different kernel sizes and downsampling methods to understand how varying receptive field sizes and spatial resolution affect AD classification performance on volumetric MRI. Architectural choices were guided by the cross-validation performance on the ADNI1/ADNI2 datasets, and the resulting configuration is referred to as AD-NeXt. Next, we explored the optimal pre-training method by varying pre-training datasets, decoder dimensions, and masking ratios, selecting the configuration that yielded the highest fine-tuning accuracy on both datasets. Finally, different variants of the best-performing architecture and pre-training scheme were compared against 3D-DenseNet121 [8], MedNeXt-B (which outperforms 3D-UX-Net, making a comparison with MedNeXt sufficient), MAE-pretrained ViT-B [6] [11], and various versions of 3D-ResNet [7] on ADNI1 and ADNI2. Models are evaluated based on the average four-fold accuracy and the area under the ROC curve (AUC), using the best-performing model for each fold. We use publicly available MONAI implementations and the training hyperparameters listed in Appendix. For the MedNeXt-B model, we follow Roy et al. [15] and set the base learning rate to 0.001 for a kernel size of 3 and 0.0001 for a kernel size of 5.

All experiments were conducted on an NVIDIA Titan RTX GPU with 24GB of VRAM.

3 Results & Discussion

3.1 Model Architecture

We begin with the ConvNeXt-V2 base model and replace its 2D convolutions with 3D counterparts. This differs from the MedNeXt encoder in its larger kernel size and simpler downsampling layer. Since kernel sizes and downsampling methods play a significant role, we perform an ablation study on ConvNeXt-V2, evaluating various kernel sizes and downsampling strategies, as shown in Figure 1. Kernel sizes of 3 and 7, along with average and max pooling layers, demonstrate superior accuracy and computational efficiency compared to strided convolutions. A combination of average pooling and a kernel size of 7 yields the highest average accuracy across the ADNI1 and ADNI2 datasets.

Therefore, AD-NeXt incorporates ConvNeXt-V2 3D blocks with depthwise convolutions with kernel size 7 to enlarge the receptive field, and uses average pooling between stages for smoother downsampling. The final architecture of our base model is depicted in Figure 2 (a).

Additionally, we introduce multiple variants of our model, each with different numbers of blocks and channels per stage:

- Tiny (T): B = (2, 2, 6, 2), C = (16, 32, 64, 128)
- Small (S): B = (3, 3, 9, 3), C = (32, 64, 128, 256)
- Base (B): B = (3, 3, 27, 3), C = (64, 128, 256, 512)
- Large (L): B = (3, 3, 27, 3), C = (128, 256, 512, 512)

For the AdNeXt-L variant, the last-stage channel size is limited to 512 rather than 1024 to maintain feasible GPU memory and training time. Expanding beyond 512 produced minimal accuracy improvement while substantially increasing computational demands and overfitting risk on the limited dataset.

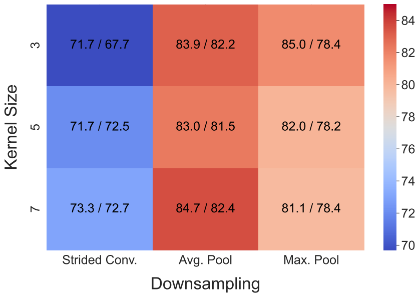

Fig. 1: Ablation on the base model varying kernel size and downsampling method. Accuracy results (%) indicate performance on ADNI2/ADNI1 respectively.

Table 2: Comparison of classification accuracy (%) between Non-sparse and Sparse Convolution in MAE pretraining

Type	ADNI1	ADNI2
Non-sparse conv.	79.82	80.83
Sparse conv.	**85.03**	**86.39**

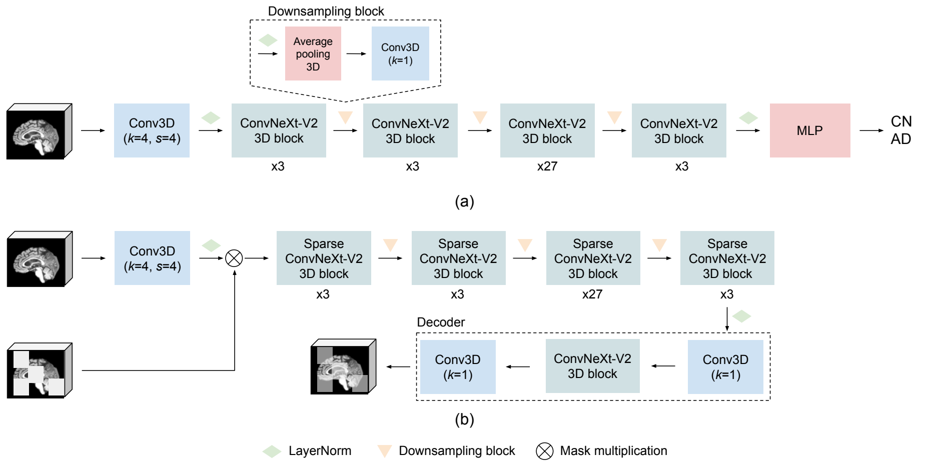

Fig. 2: Proposed AD-NeXt-B architecture for (a) fine-tuning and (b) pre-training. ConvNeXt-V2 3D block is similar to ConvNeXt-V2 block [20], except for 3D convolutional layers instead of 2D. The loss during pre-training is calculated only on the masked region.

3.2 Pre-training

During pre-training, we follow Woo *et al.* [20], where we employ a single, lightweight ConvNeXt block decoder and integrate sparse convolutions from the Minkowski Engine [3] in the encoder, as illustrated in Figure 2 (b). The efficiency of sparse convolutions is demonstrated in Table 2. Finally, we transfer the weights from the last checkpoint to the encoder for subsequent fine-tuning on downstream tasks.

We pre-train our base model on IXI, HCP, and a combination of both datasets. Figure 3 presents a comparison of fine-tuning accuracies across different pre-training datasets. Interestingly, the model achieved higher accuracy after pre-training on IXI compared to HCP, despite HCP being nearly twice the size of IXI. This discrepancy may be attributed to the heterogeneity of IXI, which includes MRI scans acquired at 1.5T and 3T field strengths, whereas HCP consists solely of 3T scans. Therefore, pre-training on the IXI dataset improves classification accuracy on ADNI1, which has the same magnet strength. This highlights the importance of diversity in pre-training datasets. However, the size of the pre-training dataset remains a crucial factor, as combining both datasets results in higher fine-tuning accuracy.

To further investigate the effect of pre-training, we compute the pairwise cosine distance between features from the first fold of ADNI1 and ADNI2, visualizing the mean and standard deviation across all data points in Figure 4. Specifically, we extract activation tensors of shape $C \times H \times W \times D$ across layers and compute the pairwise cosine distance between C features of shape $H \times W \times D$. Each feature map is flattened into a vector $X \in \mathbb{R}^{H \cdot W \cdot D}$, and the pairwise cosine

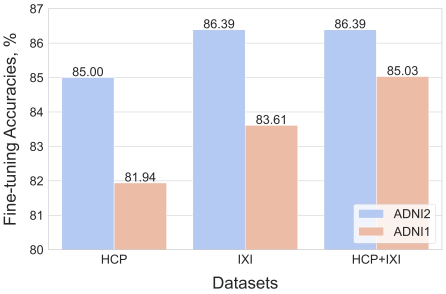

Fig. 3: Pre-training dataset effect on the downstream finetuning. A masking ratio of 60% and a decoder dimension of 128 were used.

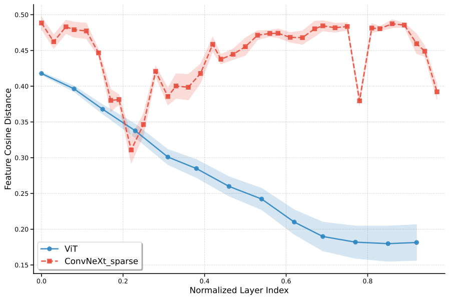

Fig. 4: Pairwise cosine distance between features. As the two models have different number of layers, x-axis depicts the normalized layer index.

distance is computed as: $\frac{1}{C^2}\sum_i^C\sum_j^C\frac{1-cos(X_i,X_j)}{2}$. Compared to ViT-MAE [11], after pre-training AD-NeXt model exhibits greater feature diversity, particularly in deeper layers, which enhances performance in downstream fine-tuning.

Additionally, we conduct ablation studies on the dimension of the decoder block and masking ratio. A key consideration for the masking ratio is the trade-off between learning rich representations and retaining sufficient spatial information. A masking ratio of 60% applied to 32×32 patches from the original volume from the original volume yields the highest accuracy for ADNI1 and ADNI2 datasets (Figure 5), while excessively high masking degrades performance. Another critical design choice is the decoder dimension. As shown in Table 3, the best performance is achieved when the decoder dimension is set to 128. Particularly in ADNI1, decoder dimension of 64 limits the model's expressiveness, whereas large dimensions (e.g., 512 and 768) introduce redundancy and lead to suboptimal performance. Therefore, for the pre-training of other AD-NeXt variations, we adopt this decoder dimension, masking ratio, and dataset combination.

3.3 Fine-tuning

Table 4 compares the classification accuracies and AUC values of models on the ADNI1 and ADNI2 datasets either trained from scratch or fine-tuned from pre-trained weights. The analysis of only AD-NeXt models shows that fine-tuning pre-trained checkpoints consistently improves performance relative to training from scratch, highlighting the effectiveness of pre-trained.

When trained from scratch, larger AD-NeXt variants generally achieve higher accuracy than smaller ones, suggesting that increased model capacity contributes to better performance. However, this trend does not persist when fine-tuning from pre-trained weights, as AD-NeXt-B outperforms its larger counterpart (AD-NeXt-L). This suggests that AD-NeXt-L may overfit due to limited pre-training

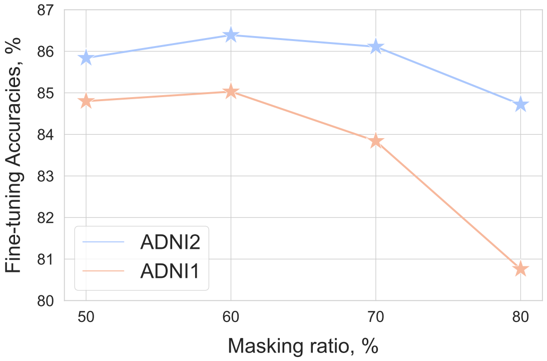

Fig. 5: Effect of masking ratio

Table 3: Effect of different decoder dimensions on classification accuracy (%)

dim	**ADNI2**	**ADNI1**
64	84.72	82.42
128	**86.39**	**85.03**
256	85.28	83.13
512	87.22	82.19
768	85.84	83.13

data diversity, leading to poorer generalization. Additional experiments with fine-tuning hyperparameters, such as learning rates and weight decay, may be necessary to fully leverage the capacity of larger models.

A notable observation is that the compact AD-NeXt-T and AD-NeXt-S trained from scratch, which has substantially fewer parameters and MACs, outperform more computationally intensive MedNeXt-B encoder. The reduced performance of MedNeXt is likely linked to its downsampling method, which creates a computational bottleneck. In contrast, AD-NeXt models use a single-layer average pooling combined with a 1×1 convolution, allowing more processing blocks which contributes to their superior performance. Additionally, pre-trained tiny and small AD-NeXt variants outperform conventional CNNs like ResNet10 and ResNet34 while remaining both lightweight and efficient.

Interestingly, DenseNet121, a smaller model than some ResNet architectures, achieves accuracy comparable to vanilla AD-NeXt-B and AD-NeXt-L on both ADNI1 and ADNI2. Nevertheless, AD-NeXt-B and AD-NeXt-L, even without pre-training, achieve higher AUC scores than ResNet101 on ADNI2 despite having slightly lower accuracy. However, after MAE-based pre-training, AD-NeXt-B achieves the top performance on both ADNI1 and ADNI2, while also requiring fewer MACs and less GPU memory than either ResNet101 or DenseNet121. Therefore, AD-NeXt-B offers the best trade-off between accuracy, computational efficiency, and memory requirements.

Finally, compared to the pre-trained ViT-B model, all pre-trained AD-NeXt variants exhibit superior performance. This can be attributed to the data-hungry nature of Vision Transformers, which typically require large-scale datasets to generalize well. Even the tiny AD-NeXt model outperform ViT-B, further underscoring the effectiveness of our architecture in handling limited data scenarios.

Table 4: Average of 4-fold classification accuracies of different models on AD datasets. **Bold** and underlined values indicates the best and second-best results, respectively.

Model	#param	MACs	ADNI1		ADNI2	
			Acc. (%)	AUC (%)	Acc. (%)	AUC (%)
MedNeXt-B (kernel=3)	2.64M	56.6G	66.99	67.19	69.44	71.58
MedNeXt-B (kernel=5)	2.78M	75.13G	70.08	71.75	71.11	70.02
DenseNet121	11.24M	43.40G	82.42	86.66	84.72	87.08
ResNet10	14.36M	143.5G	75.55	80.00	78.61	83.11
ResNet34	63.47M	433.87G	77.92	81.34	80.28	84.94
ResNet101	85.21M	487.43G	84.80	**89.56**	85.84	88.30
ViT-B (pre-trained)	88.60M	45.27G	79.58	83.41	78.33	80.97
AD-NeXt-T	0.76M	0.90G	77.43	80.74	80.56	79.82
AD-NeXt-T (pre-trained)	0.76M	0.90G	80.52	83.78	82.22	86.37
AD-NeXt-S	3.63M	3.72G	80.29	83.41	83.61	87.36
AD-NeXt-S (pre-trained)	3.63M	3.72G	80.99	84.35	84.17	87.48
AD-NeXt-B	24.36M	16.79G	82.42	86.47	84.72	88.73
AD-NeXt-B (pre-trained)	24.36M	16.79G	**85.03**	89.25	**86.39**	**89.63**
AD-NeXt-L	71.25M	57.40G	82.42	86.57	85.28	89.19
AD-NeXt-L (pre-trained)	71.25M	57.40G	82.90	88.86	85.83	88.65

4 Conclusion

In this study, we proposed AD-NeXt, an efficient architecture for MRI-based Alzheimer's disease (AD) classification, and evaluated its performance on the ADNI datasets. Our model integrates 3D depthwise convolutions, optimized downsampling with average pooling, and Masked Autoencoder (MAE) pre-training with sparse convolutions. The pre-trained AD-NeXt models consistently outperformed baseline architectures, including ResNets, DenseNet121, and ViT-B, achieving higher accuracy and AUC scores with significantly lower computational requirements.

These results demonstrate that AD-NeXt, combined with MAE pre-training, offers a competitive and computationally efficient solution for 3D medical image analysis. However, to further enhance its robustness and generalizability, future work will focus on evaluating AD-NeXt on more complex vision tasks such as semantic segmentation and exploring alternative pre-training strategies for different modalities.

References

1. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
2. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
3. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
4. 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)
5. Hazarika, R.A., Kandar, D., Maji, A.K.: An experimental analysis of different deep learning based models for alzheimer's disease classification using brain magnetic resonance images. Journal of King Saud University-Computer and Information Sciences **34**(10), 8576–8598 (2022)
6. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
7. 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)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
9. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K.H., Kickingereder, P.: Automated brain extraction of multisequence MRI using artificial neural networks. Human Brain Mapping **40**(17), 4952–4964 (aug 2019)
10. Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., and, M.W.W.: The alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging **27**(4), 685–691 (2008)
11. Kunanbayev, K., Shen, V., Kim, D.S.: Training vit with limited data for alzheimer's disease classification: An empirical study. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 334–343. Springer (2024)
12. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022)
13. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)

14. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
15. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
16. Sovrasov, V.: ptflops: a flops counting tool for neural networks in pytorch framework (2018), `https://github.com/sovrasov/flops-counter.pytorch`
17. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: Improved n3 bias correction. IEEE Transactions on Medical Imaging **29**(6), 1310–1320 (jun 2010)
18. Van Essen, D., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., Della Penna, S., Feinberg, D., Glasser, M., Harel, N., Heath, A., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S., Prior, F., Schlaggar, B., Smith, S., Snyder, A., Xu, J., Yacoub, E.: The Human Connectome Project: A data acquisition perspective. NeuroImage **62**(4), 2222–2231 (2012), Connectivity
19. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
20. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023)
21. Ye, R., Liu, F., Zhang, L.: 3d depthwise convolution: Reducing model parameters in 3d vision tasks. In: Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019, Proceedings 32. pp. 186–199. Springer (2019)

5 Appendix

5.1 Training Hyperparameters

Table 5: Fine-tuning Hyperparameters

Hyperparameter	**Value**
Optimizer	AdamW
Base LR	$1e-5$
Weight Decay	$1e-4$
LR Schedule	cosine decay [14]
Training Epochs	150
Warmup Epochs	0
Batch Size	4

Table 6: Pre-training Hyperparameters

Hyperparameter	**Value**
Optimizer	AdamW
Base LR	$1e-4$
Weight Decay	$1e-4$
LR Schedule	cosine decay
Training Epochs	600
Warmup Epochs	40
Batch Size	4