M3T: three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer

Jinseong Jang Dosik Hwang*
School of Electrical and Electronic Engineering, Yonsei University

Abstract

In this study, we propose a three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer (M3T) network to classify Alzheimer's disease (AD) in 3D MRI images. The proposed network synergically combines 3D CNN, 2D CNN, and Transformer for accurate AD classification. The 3D CNN is used to perform natively 3D representation learning, while 2D CNN is used to utilize the pre-trained weights on large 2D databases and 2D representation learning. It is possible to efficiently extract the locality information for AD-related abnormalities in the local brain using CNN networks with inductive bias. The transformer network is also used to obtain attention relationships among multi-plane (axial, coronal, and sagittal) and multi-slice images after CNN. It is also possible to learn the abnormalities distributed over the wider region in the brain using the transformer without inductive bias. In this experiment, we used a training dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) which contains a total of 4,786 3D T1-weighted MRI images. For the validation data, we used dataset from three different institutions: The Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL), The Open Access Series of Imaging Studies (OASIS), and some set of ADNI data independent from the training dataset. Our proposed M3T is compared to conventional 3D classification networks based on an area under the curve (AUC) and classification accuracy for AD classification. This study represents that the proposed network M3T achieved the highest performance in multi-institutional validation database, and demonstrates the feasibility of the method to efficiently combine CNN and Transformer for 3D medical images.

1. Introduction

Convolutional Neural Networks (CNN) have been established with a dominant performance in the computer vision field [35]. They have showed high feasibilities in various computer vision tasks such as image classification [26, 28, 35, 65], object detection [38, 53, 54], and semantic segmentation [8, 40, 55]. In addition, these CNN-based architectures have been widely applied to the medical image analysis field [39] in the various modalities such as X-ray [27], CT [24], MRI [20, 30], and Ultrasound [12], and in various dimension signals from 2D to 3D medical images [4, 74]. Especially, to analyze 3D medical images, various approaches have been established based on 2D and 3D CNN networks [4, 47, 56, 74]. The 2D-based methods have advantages from a pre-trained model using large-scale 2D natural images, while the 2D representation learning has a disadvantage for analysis of 3D image contexts [47, 56, 74]. On the other hand, The 3D-based methods can learn natively 3D representations [13, 44, 57]. However, there are few publicly available 3D databases for pretraining [61, 62]. Furthermore, the 3D model has lacks of ability to build deep layers because it requires large parameters and computa-

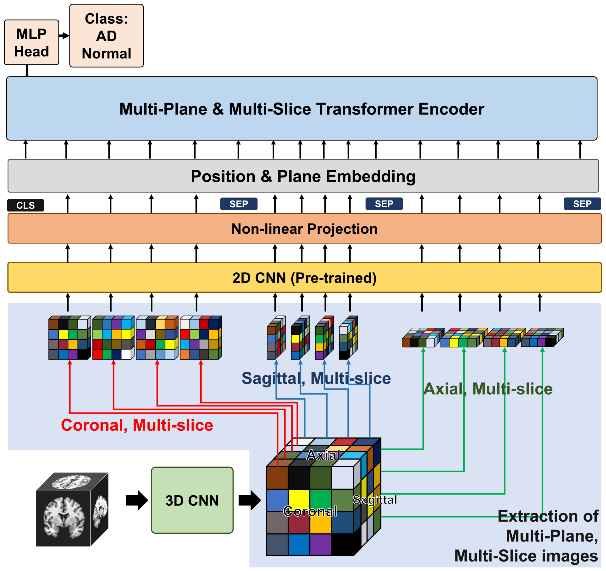

Figure 1. The overall framework of three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer network (**M3T**)

[*] Corresponding author.

tion costs [64, 72]. There are trade-offs between 2D and 3D representation learning on 3D medical images: these researches select either 2D or 3D CNN models [74].

Meanwhile, transformer networks have been widely used not only in natural language processing [16, 69] but also in computer vision processing [5, 18, 32]. These networks have a wider receptive field, which can cover a large area of images and grow linearly with the depth of the network, while the convolution-based networks have a limited receptive field. More recently, Vision Transformer (ViT) [18] which consists of a pure-transformer-based architecture could achieve reasonable performance on image classification. Furthermore, ViT achieves comparable results to conventional CNN-based methods using very large-scale databases, indicating that the transformer model is competitive with the other state-of-the-art techniques. However, when the models are trained with smaller dataset, the CNN-based method tends to show higher accuracy. This indicates that the pure-transformer-based architectures struggle to learn meaningful representations when trained on small datasets due to the low abilities of inductive biases possessed in CNN architectures [18, 32]. Especially for 3D medical images, the number of datasets is relatively lower than those of other domains because of hardly accessibilities by ethical issues [61, 62], large computational costs by high dimensionalities [64], expensive annotation, and severe class-imbalance problems [72]. Therefore, the pure-transformer-based method has not been yet widely used in analyzing 3D medical images.

In fact, there are some trade-offs between CNN and Transformer: CNN's strong inductive biases and localities to achieve high performance even with minimal data, yet these biases may limit the CNN when there are high dimensional data to cover with the low receptive field [32, 79]. On the other hand, a transformer with minimal inductive biases, which can prove to limit in small datasets, but the bias enables the architecture to cover a large area with a high receptive field [14, 18, 71]. More recently, the hybrid network combining CNN and transformer has been researched to take advantage of both methods and achieved more competitive performance compared to conventional methods [14, 71, 79]. However, these hybrid networks only combine 2D CNN and transformer for 2D images, while our method combines 2D, 3D CNN and transformer for 3D medical images.

Alzheimer's Disease (AD) is progressive neurological illness that causes memory loss and makes it difficult to communicate and perform daily tasks like walking and speaking [43]. The progression of AD often involves structural changes such as cerebral cortex atrophy, ventricle area enlargement and hippocampus volume shrinkage [1, 25]. Fig. 2 shows the brain image of normal control and AD patients. Therefore, 3D MRI images has been widely used to analyze AD-related abnormalities [63]. However, it can be challenging for doctors to analyze large and complex MRI images and to extract important information manually. Moreover, due to various inter- or intra-operator variability issues, manual analysis of brain 3D MRI is time-consuming and vulnerable to misdiagnosis [15].

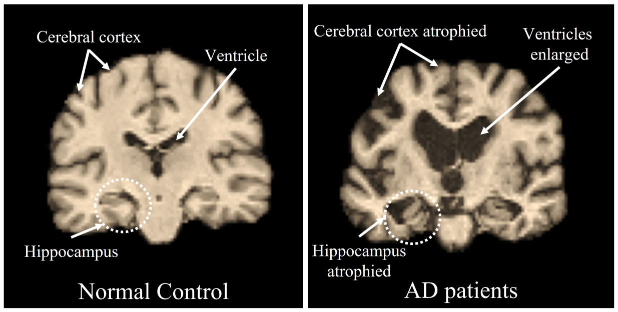

Figure 2. Comparison of a normal control brain (left) and structural changes by degeneration from severe Alzheimer's disease (right)

The atrophy of cerebral cortex of AD patients occurs in the cortex distributed throughout the brain. So, transformer architecture with a wide receptive field is suitable to detect this cortex change. On the other hand, enlargement of ventricle area and hippocampus shrinkage occur in local area of brain. A CNN network with inductive bias is suitable for these local hippocampal changes. Accordingly, we used a hybrid network combining CNN and transformer networks in this study. Furthermore, transformer network can analyze various range relationship from adjacent to far away images because it has permutation -invariant property [18].

In this study, we propose a three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer (**M3T**) network to analyze AD in 3D MRI images. Our goal is to classify Alzheimer's disease (AD) with normal control (NC) in the 3D MRI images. The overall architecture of the proposed M3T model is shown in Fig. 1. The main contributions of this study are as follows.

First, our proposed M3T successfully combines CNN and transformer architectures for 3-dimensional image classification. CNN architecture with inductive bias enables our network to efficiently analyze local features related to abnormalities of AD. The transformer with a large receptive field efficiently combines multi-plane (coronal, sagittal, and axial) and multi-slice tokens from CNN and captures a long-range relationship in 3D MRI images. M3T achieves higher performance compared with pure CNN and transformer methods.

Second, we efficiently combined 2D and 3D CNN architectures for 3D MRI images using hybrid networks and multi-plane, -slice feature extraction. Using 3D CNN, 3D representation features can be obtained to analyze 3D AD-related abnormalities. In addition, cross-slice and cross-plane 2D features used in the 2D CNN process can be extracted from the 3D images. Using 2D CNN, we use the large-database pre-trained network, which ensures stable training with a small number of medical images. For these reasons, our M3T obtain higher performance compared to the approaches that do not combined 2D or 3D CNN networks.

Third, we visualize the activated area in 3D MRI images the transformer interpretability methods [7]. These results provide explanation and interpretation of AD-related abnormalities in 3D MRI images. Furthermore, the activated area shows where the network focused on AD-related features. From these visualization results, the regions analyzed by our proposed largely coincide with the regions mainly shrunk by Alzheimer's disease.

2. Related Works

2.1. Transformer for Computer Vision

With successful application of transformer in the NLP field [16, 69], many studies have been established to lead to the transformer networks for vision tasks. The studies such as ViT [18] and DeiT [66] for image classification, DETR [5] for objection detection, ViViT [2] and VTN [46] for video analysis, and transformer-based segmentation [68]. Especially, ViT solved the first problem by simply dividing the image into non-overlapping patches and using each patch as a visual token. ViT shows that the transformer model trained on a large datasets can achieve very competitive performance for image analysis. However, when there is not enough training data, ViT does not achieve high performance because there is very low inductive bias. DeiT [66] alleviates the problem by introducing a regularization and augmentation pipeline into ImageNet-1K. In addition, the transformer methods have been studied in medical image segmentation [22, 70], 2D medical image classification [42], image denoising [41] and image reconstruction [33].

To take advantage of CNN and transformer, the hybrid networks combining both networks have been demonstrated in the computer vision field [14, 71, 79]. Through various ablation studies, the hybrid of CNN and transformer achieved the competitive performance among combinations of other networks including multi-layer perceptron in computer vision field. These results indicates that the combination of CNN and transformer with different roles can perform the vision tasks efficiently.

2.2. 3D Medical Image Analysis

The 2D CNN models have been widely used in 3D medical image analysis. The multi-plane representation methods are proposed where images from coronal, sagittal and axial planes, are treated as the three channels of 2D input [45, 51, 56]. This is empirically effective, but the weakness of the approach is that the three channels are not spatially aligned. Another approach uses the multi-slice-based methods where the three multi-slice images are regarded as the multi channels in 2D inputs [4, 17, 48, 75]. In addition, there are studies using both multi-plane and multi slices [50, 78]. However, these networks use only 2D CNN which cannot consider native 3D representation features.

Instead of the 2D CNN approaches, there are many methods using 3D CNN networks for 3D medical image analysis [13, 44, 57]. Compared to the limitation of 2D CNN networks in 3D representational learning, the 3D CNN-based methods are able to learn 3D representation features. Therefore, the 3D CNN-based approaches are generally better at tasks requiring analysis such as 3D organs in medical images. However, it is very difficult to obtain large-scale universal 3D pre-training. For this reason, efficient training of 3D networks is a pain point for 3D approaches. In addition, 3D CNN has lack of ability to build deep layers because it requires large parameters and computation costs, which causes low receptive field and has low ability to analyze a large object in 3D medical images.

To overcome the limitation of both models, we combine 2D approaches that analyze multi-plane and multi-slice images with 3D CNN methods that have 3D representation learning [74]. In addition, we use a transformer network that effectively analyzes the long-range relationship to cover the multi-plane and multi-slice features.

2.3. Alzheimer's Disease Classification

There have been deep learning-based AD classification methods. 3D VGGNet [37], ResNet [31, 34, 37, 73] and densenet [58] are used to classify AD scans. In this work, some well-known baseline 2D deep architectures, such as VGGNet and ResNet , were converted to their 3D counterparts, and the classification of AD was performed using MRI data. In addition, an auto-encoder based method to derive an embedding from the input features of 3D patches is demonstrated [36]. The combination stacked recurrent neural network with 3D CNN layers are developed for AD classification using MRI data [21]. Deep 3D CNN methods also are studied using 3D medical image for AD classification. Most of the researchers have used CNN-based networks [19, 77].

3. Methods

3.1. Proposed Network: M3T

To establish our model design, we combined various deep learning structures including 3D CNN, 2D CNN, and Transformer networks. The detailed architecture of M3T is shown in Fig. 3. M3T consists of five main blocks: 1) a 3D

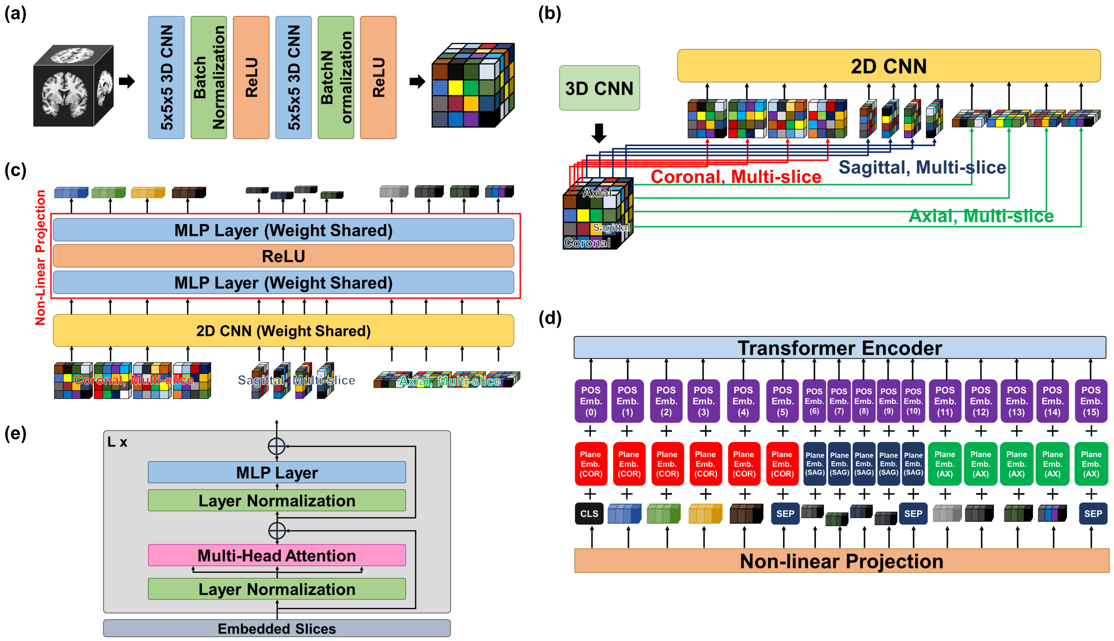

Figure 3. Detailed architecture of our proposed M3T. (a) 3D CNN model part in M3T. (b) Extraction of Multi-plane and Multi-slice images part (c) 2D CNN model and non-linear projection part in M3T. (d) Position and Plane embedding part. (e) Transformer encoder part.

CNN block to obtain natively 3D representation features, 2) extraction block of multi-plane and multi-slice tokens from 3D representation features, 3) 2D CNN to utilize the pretrained weights on a large 2D database and 2D representation features with non-linear projection network, 4) embedding block to retain position and plane information for multi-plane and -slice tokens, and 5) transformer network to obtain overall relationship among multi-plane (axial, coronal and sagittal) and multi-slice images with positional and plane embedding.

3.2. 3D Convolutional Neural Network Block

To obtain 3D representation features, we apply 3D CNN block to the MRI image $\mathbf{I} \in \mathbb{R}^{L\times W\times H}$ where image length L, width W and height H are all the same. The 3D CNN block $\mathbf{D}_{3d} : \mathbb{R}^{L\times W\times H} \rightarrow \mathbb{R}^{C_{3d}\times L\times W\times H}$ consists of two layers of $5 \times 5 \times 5$ 3D CNN layer with batch normalization and ReLU activation (C is channel number). After the 3D CNN block is applied into the input image $\mathbf{I}$, the 3D representation features $\mathbf{X}$ is calculated.

$$\mathbf{X} = \mathbf{D}_{3d}(\mathbf{I}). \tag{1}$$

The spatial size of $\mathbf{X} \in \mathbb{R}^{C_{3d}\times L\times W\times H}$ is same with the input I. Fig. 3(a) presents the detailed architecture of the 3D CNN block to obtain 3D representation features.

3.3. Extraction of Multi-plane, Multi slice images

After using 3D CNN block into the input image, the multi-plane and multi-slice image features is extracted from the 3D representation features $\mathbf{X}$. The features are calculated from the extraction operator $\mathbf{E}$. The operator consists of coronal features extractor $\mathbf{E}_{cor} : \mathbb{R}^{C_{3d}\times L\times W\times H} \rightarrow \mathbb{R}^{C_{3d}\times N\times W\times H}$, sagittal features extractor $\mathbf{E}_{sag} : \mathbb{R}^{C_{3d}\times L\times W\times H} \rightarrow \mathbb{R}^{C_{3d}\times L\times N\times H}$, and axial features extractor $\mathbf{E}_{ax} : \mathbb{R}^{C_{3d}\times L\times W\times H} \rightarrow \mathbb{R}^{C_{3d}\times L\times W\times N}$ as below Eq. (2).

$$\mathbf{E} = [\mathbf{E}_{cor}, \mathbf{E}_{sag}, \mathbf{E}_{ax}]. \tag{2}$$

Using the extractor $\mathbf{E}$, multi-plane and multi-slice features $\mathbf{S}$ are calculated from the 3D representation features $\mathbf{X}$:

$$\mathbf{S} = [\mathbf{S}_{cor}, \mathbf{S}_{sag}, \mathbf{S}_{ax}], \tag{3}$$

where the features consist multi-plane image slice feature from Eq. (3): Coronal slice feature $\mathbf{S}_{cor} \in \mathbb{R}^{C_{3d}\times N\times W\times H}$, sagittal slice feature $\mathbf{S}_{sag} \in \mathbb{R}^{C_{3d}\times L\times N\times H}$, and axial feature $\mathbf{S}_{ax} \in \mathbb{R}^{C_{3d}\times L\times W\times N}$. Because feature width, length and height are all same, concatenate and feature reshape process: $\mathbb{R}^{C_{3d}\times L\times W\times H} \rightarrow \mathbb{R}^{3N\times C_{3d}\times L\times L}$ can be applied into all the extracted features $\mathbf{S}$. Fig. 3(b) shows the detailed architecture of the extraction block to acquire the features.

3.4. 2D Convolutional Neural Network Block

The 2D CNN block process consists of two-parts: pretrained 2D CNN part and non-linear projection part. First, the weight shared 2D CNN $\mathbf{D}_{2d} : \mathbb{R}^{3N\times C_{3d}\times L\times L} \rightarrow \mathbb{R}^{3N\times C_{2d}}$ (C_{2d} is out channel size of 2D CNN) is applied to the reshaped features $\mathbf{S} \in \mathbb{R}^{3N\times C_{3d}\times L\times L}$. The 2D CNN performs global average pooling like ResNet network [26].

$$\mathbf{K} = \mathbf{D}_{3d}(\mathbf{S}), \tag{4}$$

where 2D CNN processed features $\mathbf{K} = \mathbb{R}^{3N\times C_{2d}}$. After that, we apply non-linear projection layer $\mathbf{D}_{mlp} : \mathbb{R}^{3N\times C_{2d}} \rightarrow \mathbb{R}^{3N\times d}$ widely used in various self-supervised learning for projection [10, 11, 76]. The non-linear projection consists of two layers MLP with ReLU activation between them. Using the layer, channel number C_{2d} is changed to projection dimension d. Multi-plane and multi-slice image tokens $\mathbf{T} \in \mathbb{R}^{3N\times d}$ from 2D CNN and non-linear projection layers.

$$\mathbf{T} = \mathbf{D}_{mlp}(\mathbf{K}), \tag{5}$$

where the tokens $\mathbf{T} = [\mathbf{T}_{cor}, \mathbf{T}_{sag}, \mathbf{T}_{ax}]$, coronal slice token $\mathbf{T}_{cor} \in \mathbb{R}^{N\times d}$, sagittal slice token $\mathbf{T}_{sag} \in \mathbb{R}^{N\times d}$, and axial slice token $\mathbf{T}_{ax} \in \mathbb{R}^{N\times d}$

Fig. 3(c) shows the detailed process of 2D CNN block and non-linear projection.

3.5. Position and Plane Embedding Block

After calculating the multi-plane and multi-slice image tokens, position and plane embedding tokens are added to the image tokens from non-linear projection layer, as it can be shown Fig. 3(d). First, the learnable one-dimensional position embedding tokens $\mathbf{P_{pos}}$ are applied to the embedding scheme to retain positional information. In addition, we add the plane embedding $\mathbf{P_{pln}}$ to give information indicating which plane these tokens belong to.

A learnable classification token z_{cls} is prepended to these tokens, similar to ViT class token. Plane separation tokens z_{sep} are also appended between each plane token and the end of the tokens, similar to BERT sep token. The final token used in the transformer encoder $\mathbf{Z_0} \in \mathbb{R}^{(3N+4)\times d}$ is below:

$$\begin{aligned}\mathbf{Z_0} = [&z_{cls}, T^1_{cor}, T^2_{cor}, ..., T^N_{cor}, z_{sep},\\ &T^1_{sag}, T^2_{sag}, ..., T^N_{sag}, z_{sep},\\ &T^1_{ax}, T^2_{ax}, ..., T^N_{ax}, z_{sep}] + \mathbf{P_{pos}} + \mathbf{P_{pln}},\end{aligned} \tag{6}$$

where $z_{cls} \in \mathbb{R}^d$, $z_{sep} \in \mathbb{R}^d$, $\mathbf{P_{pos}} \in \mathbb{R}^{(3S+4)\times d}$, $\mathbf{P_{pln}} \in \mathbb{R}^{(3S+4)\times d}$.

3.6. Transformer Block

Fig. 3(e) shows the transformer block architectures. The image tokens from the embedding process are then passed through consisting of a sequence of $\mathbf{K}$ transformer layers. Each layer comprises of Multi-Headed Self-Attention ($\mathbf{MSA}$) [69], layer normalization ($\mathbf{LN}$), and $\mathbf{MLP}$ blocks as follows:

$$\tilde{\mathbf{Z}}_k = \mathbf{MSA}(\mathbf{LN}(\mathbf{Z}_k) + \mathbf{Z}_k \tag{7}$$

$$\mathbf{Z}_{k+1} = \mathbf{MLP}(\mathbf{LN}(\tilde{\mathbf{Z}}_k) + \tilde{\mathbf{Z}}_k \tag{8}$$

The MLP layer consists of two linear projections separated by a GELU activation function and the token-dimensionality, d, remains fixed throughout all layers, as shown in Fig. 3(e). Finally, a linear classifier is used to classify the encoded input based on the MLP head: $z^K_{cls} \in \mathbb{R}^d$. There are two final categorization classes: NC and AD.

4. Experiments

4.1. Experimental dataset

In this study, we have acquired a training dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) for the training process. The number of the total training dataset is 4,786, including 3,174 NC and 1,612 AD cases. All MR images were obtained using 1.5T or 3T MR system, and 3D T1-weighted MRI images that have various matrix sizes, voxel spacing, and field of view (FOV). During the training, 20% of the total training dataset was used for the validation dataset, which was a patient-based random split.

To evaluate the performance of the various deep learning models, test datasets were acquired from three institutions: ADNI, Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL), and The Open Access Series of Imaging Studies (OASIS). Especially, the test database from ADNI was totally separated from the training dataset. The ADNI test dataset includes a total of 751 cases which consist of 509 NC and 242 AD cases. The AIBL dataset contains a total of 817 cases which consist of 697 NC and 120 AD cases. The OASIS dataset consists of a total of 509 cases which consist of 323 NC and 206 as shown. The dataset from the three institutions requires an institutional approval process. Although they were collected with approval from the Institutional Review Board, all the databases should not be shared without permission and only be used by authorized researchers for research purposes.

4.2. Implementation details

We apply the same data pre-processing to normalize and standardize MR images from a multi-institutional database. First, we used N4 algorithm [67] to correct the intensity inhomogeneity. Next, skull stripping algorithm was performed using HD-BET network [29]. Then, we resized the images to have the same voxel spacing ($1.75mm \times 1.75mm \times 1.75mm$) and matrix size ($128 \times 128 \times 128$). Lastly, we normalized image intensities of all the voxels using the zero-mean unit-variance method.

We applied 3D CNN block to the the pre-processed input data. The 3D CNN took of size $128 \times 128 \times 128$ and con-

Model name	Params	ADNI		AIBL		OASIS	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
3D ResNet50	46.23M	0.9226	0.8868	0.8589	0.9106	0.8175	0.7996
3D ResNet50+Transformer	51.65M	0.9351	0.9161	0.8698	0.9094	0.8504	0.8053
3D ResNet101	85.33M	0.9355	0.8908	0.8832	0.9168	0.8623	0.8147
3D ResNet101+Transformer	90.75M	0.9528	0.9148	0.9012	0.9143	0.8652	0.8185
3D ResNet152	117.54M	0.9356	0.8961	0.8718	0.8996	0.8404	0.8015
3D ResNet152+Transformer	122.96M	0.9387	0.9134	0.9071	0.9155	0.8385	0.8015
3D DenseNet201	25.60M	0.9530	0.9201	0.8975	0.9241	0.8604	0.8204
3D DenseNet201+Trasnformer	30.95M	0.9435	0.9041	0.9179	0.9253	0.8451	0.8223
3D ViT	33.87M	0.8851	0.8349	0.8173	0.8739	0.8379	0.7996
MRNet	24.75M	0.9405	0.9014	0.9050	0.9155	0.8538	0.7996
I3D	12.30M	0.9276	0.8921	0.8639	0.8984	0.8457	0.8034
MedicalNet	46.19M	0.9522	0.9081	0.9016	0.8984	0.8861	0.8261
FCNlinksCNN	12.84M	0.9489	0.9081	0.9104	0.9155	0.8495	0.8015
M3T (Ours)	**29.12M**	**0.9634**	**0.9321**	**0.9258**	**0.9327**	**0.8961**	**0.8526**

Table 1. Comparison with various 3D classification networks on multi-institutional Alzheimer's disease database.

volved them into 3D representation features with 32 channels. In addition, we used ImageNet pre-trained ResNet50 network [26] for 2D CNN block. The number of features in first MLP layer is 512, and the number of final features is 256. The number 256 is same with projection dimension (attention dimension) d used in the transformer. The number of transformer layers is 8. The hidden size and MLP size are 768, and the number of heads = 8.

We implemented M3T using a Pytorch library [49]. M3T was trained using an Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 50 epochs with a learning rate of 0.00005, and the batch size is 4. For binary classification (AD and NC), we used binary cross-entropy loss. The training took approximately 20h using NVIDIA NVIDIA TITAN RTX GPU.

Two metrics including an area under curve (AUC) and accuracy were used to quantitatively evaluate the performance of classification algorithm.

4.3. Comparison study results

We compared M3T with conventional 3D classification methods based on 3D ResNet (50, 101, 152) [26], 3D DenseNet121 [28] because they have been widely used for AD classification [19, 31, 34, 37, 58, 73, 77]. We also compared I3D [6] , MRNet [4], MedicalNet [9], and FCNlinksCNN [52]. The MRNet used in this experiment was based on 2D ResNet50 because it has higher performance than using AlexNet. I3D and MedicalNet used weights trained on Kinetics and 23 medical databases, respectively. The other networks did not use the pre-trained weights.

In this experiment, we added a hybrid network combining some 3D CNN networks with a transformer. The transformer used 3D feature tokens after the 3D CNN. In addition, we implemented 3D ViT [18] which is composed of pure-transformer networks. In this model, the sequence in transformer is applied extracted 3D patch embedding, which size is 16 × 16 × 16, and projection dimension is 512.

The quantitative performance is presented in Table 1 which shows AUC, Accuracy values of AD classification from multi-institutional datasets. M3T achieves the highest values of the metrics compared to the other methods. Except for DenseNet121 network, the performances of hybrid models combining CNN and transformer are also higher than plain 3D CNN models, which highlights the importance of transformer networks in classifying AD.

In addition, the 3D ViT has lower performance than that of the other algorithms. Although the network achieves high performance in the experiments using a very large database, the pure-transformer networks obtain low performance in our experiment with a small amount of data. On the other hand, Our proposed M3T using a hybrid network achieves competitive performance in the low amount of medical images.

4.4. Ablation study results

To evaluate the degree to which each block of the M3T network affects the performance, we compared the original M3T model with 3 models as follows: 1) M3T without initial 3D CNN block, 2) M3T without 2D CNN block and 3) without Transformer block. Table 2 shows the performance comparison results. Because of the number of parameters of the two-layer projection that directly converts the 2D multi-plane images into a one-dimensional vector, the total number of parameters in the 'w/o 2D CNN blocks' model is different from a value subtracting that of the 2D

Model	Params	ADNI		AIBL		OASIS	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
w/o 3D CNN blocks	28.98M	0.9515	0.9228	0.8896	0.9168	0.8750	0.8242
w/o 2D CNN blocks	29.67M	0.9236	0.8961	0.8617	0.8898	0.8340	0.8147
w/o Transformer blocks	24.91M	0.9445	0.9081	0.9052	0.9131	0.8727	0.8336
M3T (Ours)	**29.12M**	**0.9634**	**0.9321**	**0.9258**	**0.9327**	**0.8961**	**0.8526**

Table 2. Quantitative comparison of AD classification using 4 different models of M3T to evaluate the degree to which each block of the M3T network affects the performance.

Data extraction scheme	Params	ADNI		AIBL		OASIS	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
Single-Slice, Multi-Plane	29.09M	0.9086	0.8588	0.8640	0.8152	0.8152	0.7731
Multi-Slice, Coronal	29.11M	0.9496	0.9134	0.9101	0.9094	0.8692	0.8185
Multi-Slice, Sagittal	29.11M	0.9295	0.8975	0.8776	0.9033	0.8691	0.8223
Multi-Slice, Axial	29.11M	0.9541	0.9161	0.8617	0.9070	0.8760	0.8336
Multi-Slice, Multi-Plane	**29.12M**	**0.9634**	**0.9321**	**0.9258**	**0.9327**	**0.8961**	**0.8526**

Table 3. Quantitative comparison of AD classification using 5 different models of M3T to evaluate multi-slice and multi-plane image extraction.

CNN from that of the M3T. It represents that the hybrid model combining all blocks archives the best performance. Although the order of performance is M3T without initial 3D CNN block, without transformer block, and without 2D CNN block, it shows that all the blocks are important to classify AD in 3D MRI images.

Next, to analyze the importance of the multi-plane and multi-slice features-based method, we compared the M3T with 4 models as follows: 1) Multi-plane but single-slice, 2) Multi-slice but only coronal plane, 3) only sagittal plane and 4) only axial plane. In Table 3, M3T which uses Multi-plane and Multi-slice has the highest accuracy results compared to other models. It represents that the multi-plane and multi-slice extraction is very important to analyze the 3D MRI images. In addition, in single plane experiment cases, the axial and coronal-based model has higher performance than the sagittal-based model. Considering clinicians mainly analyze the ventricle enlargement in the axial or coronal planes, and hippocampus atrophy in the coronal plane [59,60], M3T has different abilities to analyze in each plane. However, when considering the highest performance of the multi-plane-based model, we can observe of the importance to use all of the three planes in classifying 3D MRI images.

4.5. Visualization results

We visualize the activated area of our M3T network based on transformer interpretability technique [7]. Fig. 4 shows an AD-related activation map in 3D MRI images of multi-institutional datasets. The activated maps are mainly focused on hippocampus, ventricle, and cerebral cortex areas. Especially, the axial image of Fig. 4(d) shows that M3T mainly focuses on the severely contracted cortex region in the circle annotated area. It can be seen that M3T efficiently analyzes brain structural changes that occur mainly in AD patients. On the other hand, Fig. 5 shows that the heatmap areas are widely distributed on the brain. It means that AD-related abnormalities throughout the brain can be analyzed by our proposed model. The wide activated areas are one of the advantages of the transformer networks with a high receptive field.

Fig. 6 shows the average activation map of all AD cases in 3D MRI template. The heatmap focuses mainly on the hippocampus area of the coronal plane, and the ventricle region of the axial domain. Interestingly, the right hippocampus is more focused than the left hippocampus in Fig. 6, it was studied that shrinkage of the right hippocampus occurred more in the brain of AD patients [3,23]. It shows that M3T successfully focuses on AD-related structural changes in the actual brain.

5. Conclusion

In this paper, we proposed a 3D medical image classification method, called M3T, that uses a multi-plane and multi-slice transformer for Alzheimer's disease analysis. Our proposed method combines 2D CNN, 3D CNN, and transformer networks. Experimental results show that our proposed M3T achieves higher performance compared to conventional 3D image classification network in multi-institutional test datasets. The visualization results using the transformer interpretability technique also show that M3T can visualize the AD-related regions of 3D MRI images, and the activated areas are strongly correlated with AD-related region studies in clinical research.

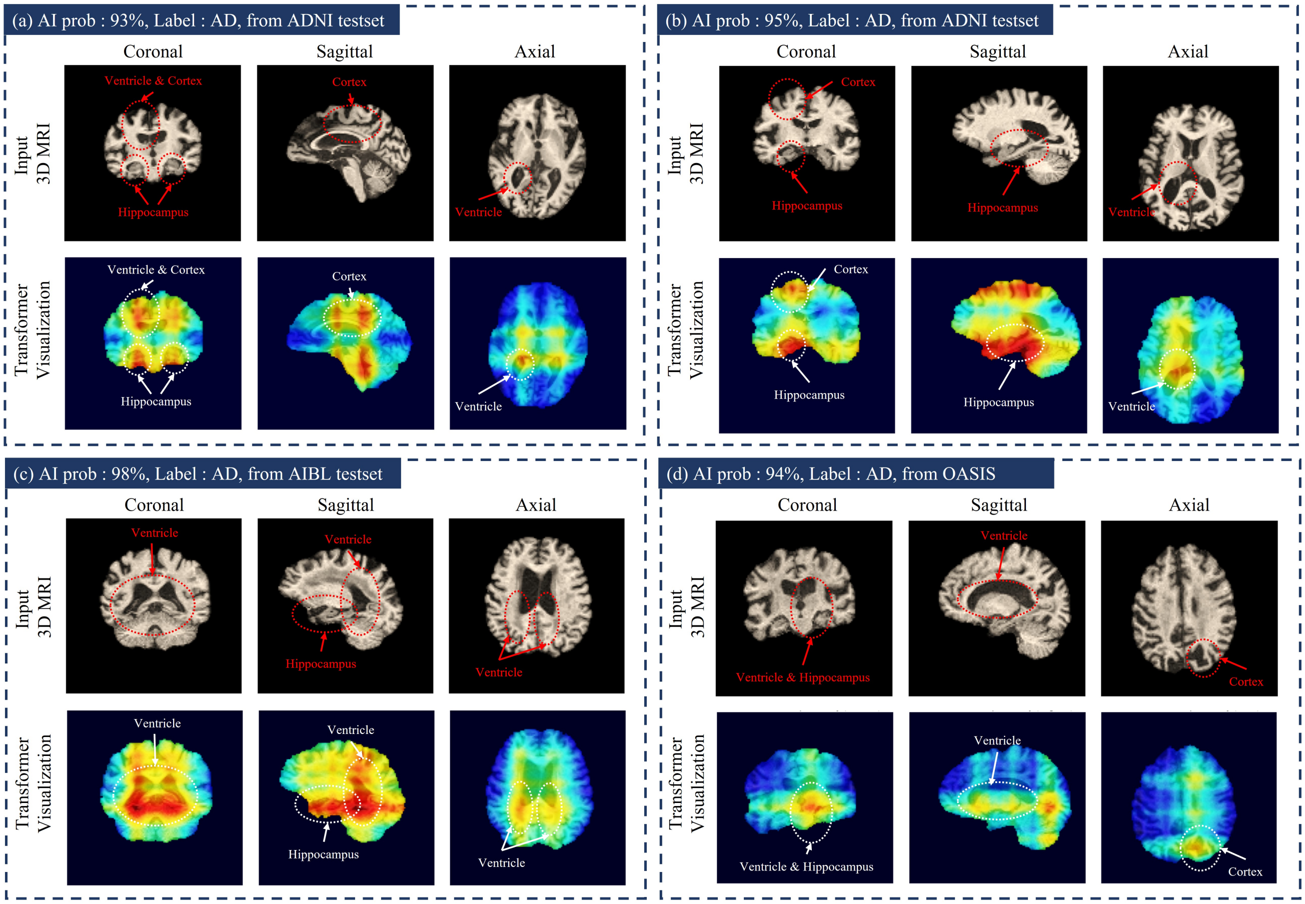
(a) AI prob : 93%, Label : AD, from ADNI testset
Coronal
Sagittal
Axial
Input 3D MRI
Transformer Visualization
Ventricle & Cortex
Hippocampus
Cortex
Ventricle
(b) AI prob : 95%, Label : AD, from ADNI testset
Coronal
Sagittal
Axial
Input 3D MRI
Transformer Visualization
Cortex
Hippocampus
Ventricle
(c) AI prob : 98%, Label : AD, from AIBL testset
Coronal
Sagittal
Axial
Input 3D MRI
Transformer Visualization
Ventricle
Hippocampus
(d) AI prob : 94%, Label : AD, from OASIS
Coronal
Sagittal
Axial
Input 3D MRI
Transformer Visualization
Ventricle & Hippocampus
Ventricle
Cortex

Figure 4. AD-related visualization map results using transformer interpretability. The results contain multi-plane images and transformer visualization results from ADNI test datasets (a), (b), AIBL dataset (b), and OASIS dataset (c). The heatmap scale is jet colormap that red color is close to one (high activated value) and blue close to zero (low activated value).

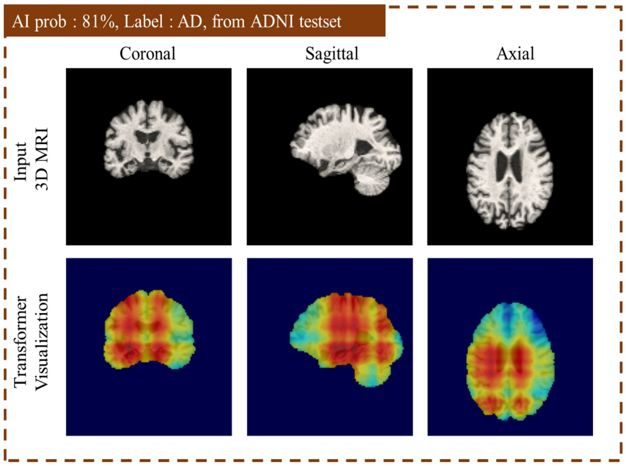
AI prob : 81%, Label : AD, from ADNI testset
Coronal
Sagittal
Axial
Input 3D MRI
Transformer Visualization

Figure 5. AD-related visualization case in which our network analyzes at the whole area rather than at a local organ of the brain.

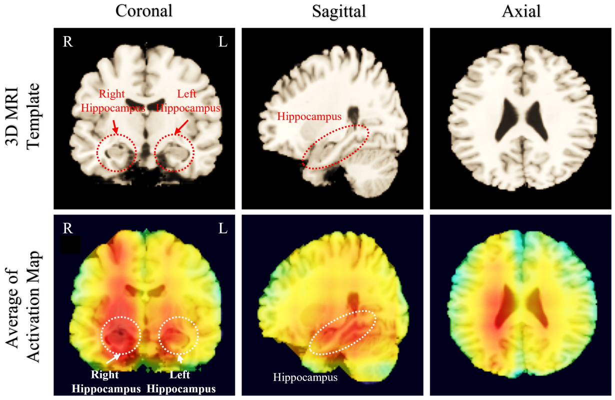

Figure 6. 3D MRI template images (first row) and average activation visualization map of all the AD cases (second row).

6. Acknowledgements

This research was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TF2103-01.

References

[1] Liana G Apostolova, Amity E Green, Sona Babakchanian, Kristy S Hwang, Yi-Yu Chou, Arthur W Toga, and Paul M Thompson. Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment and alzheimer's disease. *Alzheimer disease and associated disorders*, 26(1):17, 2012. 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 3

[3] Josephine Barnes, Rachael I Scahill, Jonathan M Schott, Chris Frost, Martin N Rossor, and Nick C Fox. Does alzheimer's disease affect hippocampal asymmetry? evidence from a cross-sectional and longitudinal volumetric mri study. *Dementia and geriatric cognitive disorders*, 19(5-6):338–344, 2005. 7

[4] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine*, 15(11):e1002699, 2018. 1, 3, 6

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 3

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 3, 7

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[9] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 6

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 5

[12] Tsung-Chen Chiang, Yao-Sian Huang, Rong-Tai Chen, Chiun-Sheng Huang, and Ruey-Feng Chang. Tumor detection in automated breast ultrasound using 3-d cnn and prioritized candidate aggregation. *IEEE transactions on medical imaging*, 38(1):240–249, 2018. 1

[13] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 1, 3

[14] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021. 2, 3

[15] Ivana Despotović, Bart Goossens, and Wilfried Philips. Mri segmentation of the human brain: challenges, methods, and applications. *Computational and mathematical methods in medicine*, 2015, 2015. 2

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[17] Jia Ding, Aoxue Li, Zhiqiang Hu, and Liwei Wang. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–567. Springer, 2017. 3

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2, 3, 6

[19] Amir Ebrahimi, Suhuai Luo, and Raymond Chiong. Introducing transfer leaming to 3d resnet-18 for alzheimer's disease detection on mri images. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2020. 3, 6

[20] Taejoon Eo, Yohan Jun, Taeseong Kim, Jinseong Jang, Ho-Joon Lee, and Dosik Hwang. Kiki-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic resonance in medicine*, 80(5):2188–2201, 2018. 1

[21] Chiyu Feng, Ahmed Elazab, Peng Yang, Tianfu Wang, Feng Zhou, Huoyou Hu, Xiaohua Xiao, and Baiying Lei. Deep learning framework for alzheimer's disease diagnosis via 3d-cnn and fsbi-lstm. *IEEE Access*, 7:63605–63618, 2019. 3

[22] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 61–71. Springer, 2021. 3

[23] C Geroldi, MP Laakso, Charles DeCarli, A Beltramello, A Bianchetti, H Soininen, M Trabucchi, and Giovanni B Frisoni. Apolipoprotein e genotype and hippocampal asymmetry in alzheimer's disease: a volumetric mri study. *Journal of Neurology, Neurosurgery & Psychiatry*, 68(1):93–96, 2000. 7

[24] Harshit Gupta, Kyong Hwan Jin, Ha Q Nguyen, Michael T McCann, and Michael Unser. Cnn-based projected gradient descent for consistent ct image reconstruction. *IEEE transactions on medical imaging*, 37(6):1440–1453, 2018. 1

[25] Taylor C Harris, Rijk de Rooij, and Ellen Kuhl. The shrinking brain: cerebral atrophy following traumatic brain injury. *Annals of biomedical engineering*, 47(9):1941–1959, 2019. 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5, 6

[27] Morteza Heidari, Seyedehnafiseh Mirniaharikandehei, Abolfazl Zargari Khuzani, Gopichandh Danala, Yuchen Qiu, and Bin Zheng. Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms. *International journal of medical informatics*, 144:104284, 2020. 1

[28] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1, 6

[29] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019. 5

[30] Yohan Jun, Hyungseob Shin, Taejoon Eo, and Dosik Hwang. Joint deep model-based mr image and coil sensitivity reconstruction network (joint-icnet) for fast mri. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2021. 1

[31] Hiroki Karasawa, Chien-Liang Liu, and Hayato Ohwada. Deep 3d convolutional neural network architectures for alzheimer's disease diagnosis. In *Asian conference on intelligent information and database systems*, pages 287–296. Springer, 2018. 3, 6

[32] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 2

[33] Yilmaz Korkmaz, Salman UH Dar, Mahmut Yurt, Muzaffer Özbey, and Tolga Çukur. Unsupervised mri reconstruction via zero-shot learned adversarial transformers. *arXiv preprint arXiv:2105.08059*, 2021. 3

[34] Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 835–838. IEEE, 2017. 3, 6

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[36] KR Kruthika, HD Maheshappa, Alzheimer's Disease Neuroimaging Initiative, et al. Cbir system using capsule networks and 3d cnn for alzheimer's disease diagnosis. *Informatics in Medicine Unlocked*, 14:59–68, 2019. 3

[37] Qi Li and Mary Qu Yang. Comparison of machine learning approaches for enhancing alzheimer's disease classification. *PeerJ*, 9:e10549, 2021. 3, 6

[38] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[39] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[41] Achleshwar Luthra, Harsh Sulakhe, Tanish Mittal, Abhishek Iyer, and Santosh Yadav. Eformer: Edge enhancement based transformer for medical image denoising. *arXiv preprint arXiv:2109.08044*, 2021. 3

[42] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images? *arXiv preprint arXiv:2108.09038*, 2021. 3

[43] Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):263–269, 2011. 2

[44] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 1, 3

[45] Pim Moeskops, Jelmer M Wolterink, Bas HM van der Velden, Kenneth GA Gilhuijs, Tim Leiner, Max A Viergever, and Ivana Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 478–486. Springer, 2016. 3

[46] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021. 3

[47] Tianwei Ni, Lingxi Xie, Huangjie Zheng, Elliot K Fishman, and Alan L Yuille. Elastic boundary projection for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2019. 1

[48] Tianwei Ni, Lingxi Xie, Huangjie Zheng, Elliot K Fishman, and Alan L Yuille. Elastic boundary projection for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2019. 3

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An

imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 6

[50] Mathias Perslev, Erik Bjørnager Dam, Akshay Pai, and Christian Igel. One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 30–38. Springer, 2019. 3

[51] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International conference on medical image computing and computer-assisted intervention*, pages 246–253. Springer, 2013. 3

[52] Shangran Qiu, Prajakta S Joshi, Matthew I Miller, Chonghua Xue, Xiao Zhou, Cody Karjadi, Gary H Chang, Anant S Joshi, Brigid Dwyer, Shuhan Zhu, et al. Development and validation of an interpretable deep learning framework for alzheimer's disease classification. *Brain*, 143(6):1920–1933, 2020. 6

[53] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[56] Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M Summers. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In *International conference on medical image computing and computer-assisted intervention*, pages 520–527. Springer, 2014. 1, 3

[57] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. An application of cascaded 3d fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, 2018. 1, 3

[58] Juan Ruiz, Mufti Mahmud, Md Modasshir, M Shamim Kaiser, for the Alzheimer's Disease Neuroimaging Initiative, et al. 3d densenet ensemble in 4-way classification of alzheimer's disease. In *International Conference on Brain Informatics*, pages 85–96. Springer, 2020. 3, 6

[59] Philip Scheltens, Leonore J Launer, Frederik Barkhof, Henri C Weinstein, and Willem A van Gool. Visual assessment of medial temporal lobe atrophy on magnetic resonance imaging: interobserver reliability. *Journal of neurology*, 242(9):557–560, 1995. 7

[60] Philip Scheltens, Florence Pasquier, Jan GE Weerts, Frederik Barkhof, and Didier Leys. Qualitative assessment of cerebral atrophy on mri: inter-and intra-observer reproducibility in dementia and normal aging. *European neurology*, 37(2):95–99, 1997. 7

[61] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017. 1, 2

[62] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 1, 2

[63] Rebecca Smith-Bindman, Diana L Miglioretti, Eric Johnson, Choonsik Lee, Heather Spencer Feigelson, Michael Flynn, Robert T Greenlee, Randell L Kruger, Mark C Hornbrook, Douglas Roblin, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. *Jama*, 307(22):2400–2409, 2012. 2

[64] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. 2

[65] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1

[66] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[67] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010. 5

[68] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*, 2021. 3

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 3, 5

[70] Madeleine K Wyburd, Nicola K Dinsdale, Ana IL Namburete, and Mark Jenkinson. Teds-net: Enforcing diffeomorphisms in spatial transformers to guarantee topology preservation in segmentations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 250–260. Springer, 2021. 3

[71] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *arXiv preprint arXiv:2106.14881*, 2021. 2, 3

[72] Ke Yan, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M Summers. Holistic and comprehensive annotation of clinically significant findings on diverse ct images: learning from radiology reports and label ontology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8523–8532, 2019. 2

[73] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Visual explanations from deep 3d convolutional neural networks for alzheimer's disease classification. In *AMIA annual symposium proceedings*, volume 2018, page 1571. American Medical Informatics Association, 2018. 3, 6

[74] Jiancheng Yang, Xiaoyang Huang, Yi He, Jingwei Xu, Canqian Yang, Guozheng Xu, and Bingbing Ni. Reinventing 2d convolutions for 3d images. *IEEE Journal of Biomedical and Health Informatics*, 2021. 1, 2, 3

[75] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K Fishman, and Alan L Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8280–8289, 2018. 3

[76] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 5

[77] Jie Zhang, Bowen Zheng, Ang Gao, Xin Feng, Dong Liang, and Xiaojing Long. A 3d densely connected convolution neural network with connection-wise attention mechanism for alzheimer's disease classification. *Magnetic Resonance Imaging*, 78:119–126, 2021. 3, 6

[78] Long Zhang, Ruoning Song, Yuanyuan Wang, Chuang Zhu, Jun Liu, Jie Yang, and Lian Liu. Ischemic stroke lesion segmentation using multi-plane information fusion. *IEEE Access*, 8:45715–45725, 2020. 3

[79] Yucheng Zhao, Guangting Wang, Chuanxin Tang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. A battle of network structures: An empirical study of cnn, transformer, and mlp. *arXiv preprint arXiv:2108.13002*, 2021. 2, 3