LongFormer: Longitudinal Transformer for Alzheimer's Disease Classification with Structural MRIs

Qiuhui Chen, Qiang Fu, Hao Bai, Yi Hong✉
Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, 200240, China
yi.hong@sjtu.edu.cn

Abstract

Structural magnetic resonance imaging (sMRI), especially longitudinal sMRI, is often used to monitor and capture disease progression during the clinical diagnosis of Alzheimer's Disease (AD). However, current methods neglect AD's progressive nature and have mostly relied on a single image for recognizing AD. In this paper, we consider the problem of leveraging the longitudinal MRIs of a subject for AD classification. To address the challenges of missing data, data demand, and subtle changes over time in learning longitudinal 3D MRIs, we propose a novel model ***LongFormer****, which is a hybrid 3D CNN and transformer design to learn from image and longitudinal flow pairs. Our model can fully leverage all images in a dataset and effectively fuse spatiotemporal features for classification. We evaluate our model on three datasets, i.e., ADNI, OASIS, and AIBL, and compare it to eight baseline algorithms. Our proposed LongFormer achieves state-of-the-art performance in classifying AD and NC subjects from all three public datasets. Our source code is available online at https://github.com/Qybc/LongFormer.*

1. Introduction

Alzheimer's Disease (AD) is one of the most common cognitive impairment diseases suffered by older people, especially in the current aging society. Medical brain scans, like Magnetic Resonance Images (MRIs), provide a non-invasive way to capture disease pathological patterns. And structural MRI (sMRI) is recommended to be a part of clinical assessment for early diagnosis of AD [28], due to its capability of characterizing brain tissue damage or loss years before the clinical symptoms appear [6, 32]. As shown in the left column of Fig. 1, compared to that of a normal control (NC) subject, the sMRI of an AD subject typically shows enlarged ventricles and hippocampus and a shrinking cerebral cortex. However, one single MRI is probably not enough to separate these two groups of subjects correctly; like the two subjects on the right column of Fig. 1, using only one of their MRIs would lead to the wrong classification result. In clinical diagnosis, the degeneration speed inferred from the followed-up or longitudinal image scans is an important factor for recognizing AD subjects [16, 28, 31]. As the right column shown in Fig. 1, by considering two scans of a subject and calculating the flow to estimate changes over time, we can separate AD subjects from normal controls more easily. Therefore, leveraging longitudinal brain MRIs is a promising way to study AD and help in computer-aided AD diagnosis.

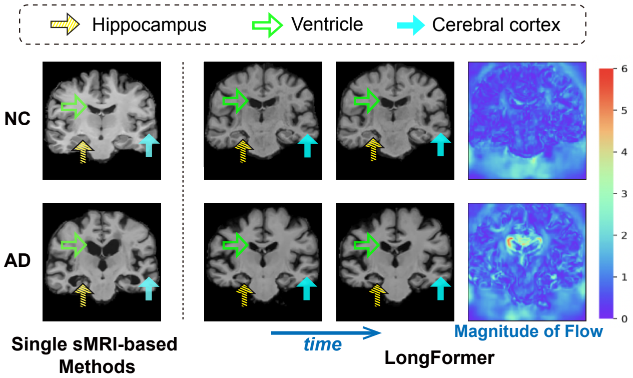

Figure 1. Our motivation for proposing *LongFormer*. Existing methods [20, 23, 40, 42] rely on the analysis of cross-sectional sMRI scans collected at a single time point, ignoring the progressive nature of AD disorder (e.g., M3T [20]). Our LongFormer considers the progression of brain atrophy from current and prior images, achieving the SOTA performance in AD classification.

In this paper, we consider learning from 3D longitudinal structural MRIs (sMRIs) to separate AD subjects from normal controls. Researchers have collected brain MRI scans of a subject at multiple time points, resulting in a longitudinal dataset for tracking the progression of AD, e.g., the well-known ADNI dataset [19], OASIS [26] and AIBL [13]. Most current methods proposed for MRI-based AD diagnosis treat their datasets as cross-sectional ones [20, 42], that is, they simply consider a single MRI for classification but ignore the degeneration progress included in the MRI sequence of a subject. However, temporal information can provide complementary supervision, solely by exploiting existing data dependency, and without requiring any additional data. Limited work has been done to fully leverage the longitudinal dataset [8]. A recent work in [16] considers two time points for studying AD but only uses 2D CNN techniques, which do not fully leverage spatial information.

Learning from longitudinal 3D image volumes faces the following three challenges: (1) The missing data. Many subjects have no scans at some time points. How to handle missing data is a non-trivial task. (2) Large memory cost and limited data size. One 3D volume has millions of voxels and needs a good amount of GPU memory for a deep-learning model. At the same time, we only have hundreds of AD or NC subjects for learning. This makes it even harder to handle a sequence of 3D volumes. (3) Small longitudinal changes and subtle subject differences. A subject's changes over time are relatively small, and the differences between AD and NC subjects are subtle to recognize.

To address these challenges, we pair each current scan of a subject with its prior image and estimate the longitudinal changes between them by computing flows via optical flow [38] or deformation fields via VoxelMorph [2]. Since we can scale the flows to normalize them, all images in the dataset can be used for learning, with no need to worry about the missing data issue. Also, because each subject can have multiple image and flow pairs, we will have thousands of input samples for learning. Besides, the pre-computed flows take some heat from the network to learn small longitudinal changes, which is experimentally demonstrated to be more effective than directly working on image pairs.

To learn from the pair of a 3D sMRI and its flow, we develop a hybrid CNN-Transformer framework named *LongFormer*. Its CNN-based embedding module reduces the input size of the following transformer and the data requirement of our model. The follow-up query-based transformer adopts a deformable attention mechanism [36, 43] to efficiently integrate the spatial and temporal features of the sMRI and its flow. Compared to eight baselines, our LongFormer achieves the state-of-the-art (SOTA) performance of classifying AD on three public datasets.

Overall, our contributions in this paper are three-fold:

- To our best knowledge, we are the first to explore an efficient vision Transformer on 3D longitudinal MRIs, which adaptively extracts spatiotemporal features from image and flow combinations for AD classification.
- We propose a novel model *LongFormer*, which provides a framework for learning from 4D data using 3D CNNs for embedding and query-based transformer with deformable cross-attention to fuse different sources of features flexibly and efficiently.
- Our LongFormer achieves the SOTA performance of classifying AD and NC subjects on three public datasets, i.e., ADNI, OASIS, and AIBL. Especially, on the largest dataset ADNI, we achieve over 93% accuracy on AD classification.

2. Related Work

Vision Transformer. Transformer is firstly proposed for the sequence-to-sequence machine translation [34] and currently becomes the basic component in most natural language processing tasks. Recently, the transformer has been successfully applied in computer vision, such as DETR [5] for object detection, SETR [41] for semantic segmentation, ViT [10] and DeiT [33] for image recognition. DETR proposes a new detection paradigm upon transformers, which simplifies object detection to a set prediction problem. Deformable DETR [43] achieves better performance by using local attention and multi-scale feature maps. To handle videos, a sequence of image frames, SeqFormer [35] adopts deformable DETR in the video instance segmentation task. SeqFormer proposes the query decomposition mechanism, splits instance queries at each frame, and then aggregates them to obtain a representation at the video level. DAT [36] proposes a deformable attention mechanism, which is to learn a set of global keys shared among visual tokens, and can be adopted as a general backbone for vision tasks. Our method adopts the deformable attention mechanism by selecting a set of learnable keys instead of global keys to learn visual combination representations.

CNN-Based AD Classification. Deep neural networks have been widely applied for AD recognition. Thanks to the publicly available large datasets, like ADNI [19], OASIS [26] and AIBL [13], training deep models for detecting AD pathology becomes possible. In 2018, a hierarchical fully convolutional network (FCN) is proposed in [25] to learn multi-scale features from both small patches and whole brain regions to perform AD diagnosis. In 2020, a multi-modality FCN with a multilayer perceptron (MLP) model is proposed in [29] to take both MRIs and associated subject attributes (e.g., age, gender) and is trained on ADNI and tested on multiple datasets. In 2021, a dual-attention multi-instance deep learning model (DA-MIDL) is proposed in [42] to identify discriminative pathological locations for AD diagnosis using sMRIs. In 2022, a three-dimensional medical image classifier is proposed in [20], using a multi-plane and multi-slice Transformer (M3T) network to classify AD using 3D MRIs. The proposed network synergically combines 3D CNN, 2D CNN, and Transformer for AD classification. A 3D Global Fourier Network (GF-Net) is proposed in [40] to utilize global frequency information that captures long-range dependency in the spatial domain. Trans-ResNet proposed in [23] integrates CNNs

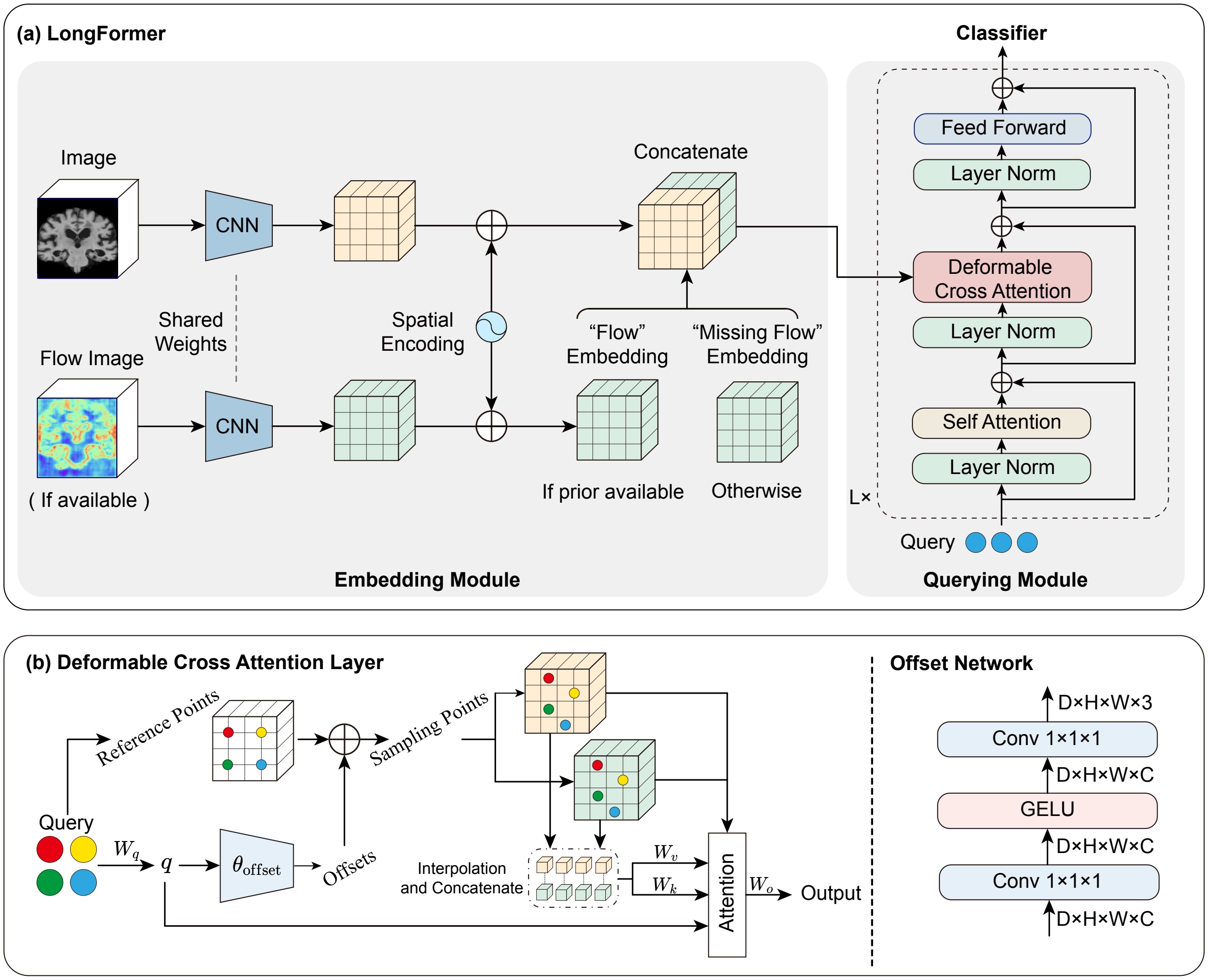

Figure 2. Illustration of our proposed LongFormer, which includes two components, i.e., the embedding module and the querying module. The deformable cross-attention layer is the core of each querying block in the querying module.

and Transformers for AD classification.

Different from these existing works that only consider cross-sectional MRIs (i.e., a single image, ignoring time information), our model integrates longitudinal changes over time, which is an important feature for recognizing AD.

Longitudinal Analysis for Image-Based AD Study. Longitudinal study tracks the changes in brain morphology of AD subjects over time. Different from the cross-sectional setting, the longitudinal one considers multiple images from the same subject scanned at different time points as a whole for analysis. This means the dimension of the input data increases while the number of training samples decreases, which brings the risk of overfitting, especially for 3D image volumes. To address this issue, researchers use a 3D CNN to extract brain image features from each MRI, and an RNN to fuse them to extract the longitudinal changes [8]. A recent work [16] extracts features from image slices of longitudinal sMRIs, i.e., the baseline and its follow-up scan, and encodes them as high-level feature representation tokens by using a transformer. Since this method takes only 2D slices and two time points, it losses both spatial and temporal information of a subject's image sequence.

After performing many experiments, we observe that directly learning from a sequence of 3D MRIs is non-trivial, due to the rapidly increased GPU memory requirement, greatly reduced input samples for learning, and the subtle longitudinal changes in the MRI sequence. Hence, we pair each image with a prior one to compute longitudinal flows between them and learn the flow for the one with the missing prior image. In this way, we can fully leverage all images in our longitudinal dataset; at the same time, we use the flow to focus on estimating the subtle longitudinal changes, which greatly reduces the learning difficulty of a network.

3. Methodology

Figure 2 presents the hybrid CNN-Transformer framework of our Longformer network, which predicts the AD classification label of a subject based on an input pair. This input pair is a combination of a current image and its associated flow which is pre-computed to measure longitudinal changes of this subject, e.g., the optical flow, or the deformation field based on image registration.

3.1. Pre-Computing Longitudinal Flow

Assume a subject has an image scan at some time point, and we treat it as the current scan $\mathbf{x}^{\text{curr}}$. To consider the longitudinal changes, e.g., within one year, we take the prior image one year ago and will face two situations, i.e., the prior image exists and we have the longitudinal pair that includes multiple images, $(\mathbf{x}^{\text{curr}}, \mathbf{x}^{\text{prior}}) \in \mathcal{D}_m$, or it does not exist and the pair includes a single image, $(\mathbf{x}^{\text{curr}}, \varnothing) \in \mathcal{D}_s$. In this way, we reformulate our longitudinal dataset and construct a new one $\mathcal{D} = \mathcal{D}_m \cup \mathcal{D}_s$.

For the image pair in the subset $\mathcal{D}_m$, we directly compute their optical flow based on [38] or deformation fields using a pre-trained VoxelMorph [2]. For the image pair in the subset $\mathcal{D}_s$, we also have two cases, i.e., one is the baseline image that has no prior image, and the other is a follow-up image that has a prior image within half year or 1.5 years ago. For the baseline image case, we leave the flow empty, which will be learned later in our LongFormer. For the follow-up image case, we compute the flow using its combination with the prior image and scale the flow using the age difference, as below:

$$I^{\text{flow}} = \frac{\Phi(I^{\text{curr}}, I^{\text{prior}})}{t^{\text{curr}} - t^{\text{prior}}}, \tag{1}$$

where Φ indicates the computation function of optical flow or deformation fields.

3.2. LongFormer Framework

As shown in Fig 2(a), our LongFormer network includes two main modules, i.e., an embedding module that extracts image and flow features based on a CNN and a querying module that fuses and learns spatiotemporal representation for AD classification based on a transformer.

The embedding module takes the input image and flow pair $(I^{\text{curr}}, I^{\text{flow}})$, where $I^{\text{curr}} \in \mathbb{R}^{D\times H\times W}$ is the 3D current image and $I^{\text{flow}} \in \mathbb{R}^{3\times D\times H\times W}$ is the associated flow in the vector form. The embedding module produces the support features $F_S \in \mathbb{R}^{C_S\times D_S\times H_S\times W_S}$ for the follow-up querying module. The querying module is then responsible for learning query representation, that is, the query features $F_Q \in \mathbb{R}^{C_Q\times N_Q}$ based on the learnable query Q and the support features F_S. Finally, the classification head produces a prediction based on the learned query features F_Q. Overall, our LongFormer can be briefly expressed as

$$\begin{aligned} F_S &= \text{Embedding}(I^{\text{curr}}, I^{\text{flow}}), \\ F_Q &= \text{Querying}\left(F_S, Q\right), \\ O &= \text{Classification Head}\left(F_Q\right), \end{aligned} \tag{2}$$

where O indicates the final prediction output.

(1) Embedding Module. Since the follow-up querying module will further extract spatiotemporal features for classification, this embedding module adopts a 3D CNN network, e.g., a ResNet [14], DenseNet [17], which serves as a backbone network to provide visual embeddings of individual images and flows. This CNN backbone is a reasonable choice because of the CNN's inductive biases [11, 27], and it helps in efficiency by reducing the input size of the following transformer-based querying module.

To further reduce the number of parameters of the backbone network, we prefer the weight-sharing technique for the two branches of the embedding module. However, the input image and flow have different sizes; therefore, we first apply a convolutional layer before sharing weights, i.e.,

$$F_S = \text{Embedding}\left(f_{a_1}(I^{\text{curr}}), f_{a_2}(I^{\text{flow}})\right). \tag{3}$$

Here, f_{a_1} and f_{a_2} are convolutional layers taking one and three input channels, respectively.

To extract support features F_S, we take the feature maps at the stage $S5$ of the backbone network, where the spatial resolution is $1/2^5 = 1/32$ of the input image. While the number of channels of the features is $C = 1024$. When no prior image is available, the flow embedding branch is not used but replaced by a learnable embedding, which is replicated across the spatial dimensions.

(2) Querying Module. The support feature F_S generated by the embedding module is a concatenation of features for a current image and its associated longitudinal flow. To fuse these two sets of feature maps, the querying module utilizes a query-based transformer as in Fig 2, which captures patch embedding interactions and aggregates them to learn a fixed-length token representation.

As shown in Fig 2(a), our querying module is a stack of L querying blocks. Each block considers the support features F_S from the embedding module with spatial position encoding and the query features F_Q from the previous layer. Using the self- and cross-attention operators, the querying block gradually learns and refines the query features. Also, to improve the flexibility of receptive fields in transformer layers, we adopt a deformable cross-attention design similar to [36] (see Fig. 2(b)), where we replace $\mathbf{q}$ to learnable queries, instead of image features itself. This transformer mechanism equipped by our querying design satisfies the need of generating flexible query features F_Q, which has large receptive fields and strong representation ability, as demonstrated by our experimental results. Details on querying blocks are included in Sec. 3.3.

(3) Classification Head. To perform classification, we can use the first query to produce a representative feature for the classifier. Another choice is using multiple uniformly distributed queries indicating different views of the subject to vote for the final prediction. In this paper, we intend to use the first one to query the category. For the loss function, since we work on the binary classification (i.e., AD and NC), we simply use the binary cross-entropy loss.

Next, let us discuss in detail the building block of our LongFormer, i.e., the querying blocks.

3.3. LongFormer's Querying Block

Each querying block takes two inputs. Take the l-th querying block for example. This block inputs the query features F_Q^{l-1} from the last layer and the support features F_S from the embedding module, and then outputs F_Q^l for the next layer. For the first querying block, we initialize F_Q^0 randomly. All L querying blocks update the query features at each iteration. In our experiments, we set L to be 6.

Each querying block consists of three layers, i.e., a self-attention layer, a deformable cross-attention layer, and a feed-forward layer (FFN). We formulate them as

$$\begin{aligned}
\hat{F}_Q^l &= \text{Self-Attention}\left(\text{LN}(F_Q^{l-1})\right) + F_Q^{l-1}, \\
\hat{F}_Q^l &= \text{Deformable Cross-Attention}\left(\text{LN}(\hat{F}_Q^l), F_S\right) + \hat{F}_Q^l, \\
F_Q^l &= \text{FFN}\left(\text{LN}(\hat{F}_Q^l)\right) + \hat{F}_Q^l.
\end{aligned} \tag{4}$$

Here, LN is a layer normalization [1] to normalize features before each attention and FFN module. The self-attention layer is a classical **qkv**-based multi-head self-attention [34], where **q**, **k**, and **v** are all from the learnable query features F_Q^{l-1}. The deformable cross-attention layer is a **qkv**-based multi-head cross-attention, where **q** is from the query features $\hat{F}_Q^l$, while **k** and **v** are from the support features F_S.

Deformable Cross-Attention Layer. This layer is responsible for integrating features from images and flows, that is, it fuses and extracts spatiotemporal features for AD classification. As illustrated in Fig. 2(b), this deformable attention layer computes cross-attention at multiple sample points $P_Q \in \mathbb{R}^{N\times 3}$, which have flexible locations within the supported features $F_S \in \mathbb{R}^{C_S\times D_S\times H_S\times W_S}$. Here, $N < D_S \times H_S \times W_S$ is the number of target positions, which can be divided into a uniform grid with $N = D_G \times H_G \times W_G$ points and are treated as the references. In particular, these reference points are uniformly located in the 3D coordinates $[(0,0,0),(D_G-1,H_G-1,W_G-1)]$. Then, we normalize them into the range $[-1,+1]$ according to the grid shape, where $(-1,-1,-1)$ indicates the top-left corner and $(+1,+1,+1)$ indicates the bottom-right corner.

To estimate the offset for each reference point, the learnable queries, i.e., $F_Q \in \mathbb{R}^{N\times d}$ (d denotes the number of channels), are projected linearly to the query tokens $q = F_Q W_q$, and then fed into a lightweight sub-network θ offset$(\cdot)$ to generate the offsets $\Delta p = \theta_{\text{offset}}(q)$. To stabilize the training process, we scale the amplitude of Δp by a pre-defined factor $s = 2$ to prevent those large offsets, i.e., $\Delta p \leftarrow s\tanh(\Delta p)$. Then the features are sampled at the locations of deformed points as keys and values, followed by a set of linear projections:

$$\begin{aligned}
&q = F_Q W_q,\ \tilde{k} = \tilde{F_S} W_k,\ \tilde{v} = \tilde{F_S} W_v \\
&\text{with } \tilde{F_S} = \phi(F_S; P_Q + \Delta p),\ \Delta p = \theta_{\text{offset}}(q),
\end{aligned} \tag{5}$$

where W_q, W_k, W_v are projection matrices, $\tilde{k}$ and $\tilde{v}$ represent deformed key and value embeddings, respectively, and $\phi(\cdot;\cdot)$ is a sampling function using trilinear interpolation.

4. Experiments

4.1. Experimental Datasets

We evaluate our model on the following three datasets.

ADNI [19]. The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset includes 5265 1.5T/3T T1-weighted structural MRI (sMRI) scans collected from 1306 subjects with visits at one or multiple time points across four ADNI phases (i.e., ADNI-1, ADNI-2, ADNI-GO, and ADNI-3). These subjects are divided into two categories, i.e., AD (including 413 subjects) and NC (including 893 subjects), according to the standard clinical criteria, e.g., the Mini-Mental State Examination (MMSE) scores and the Clinical Dementia Rating (CDR). The original structural MRI data downloaded from the ADNI website went through a series of pre-processing steps, including denoising, bias field correction, skull stripping, and affine registration to the SRI24 atlas. Then, we resample the image volumes, resulting in images of size 224 × 224 × 224, with a resolution of $1.75mm \times 1.75mm \times 1.75mm$.

OASIS [26]. Open Access Series of Imaging Studies (OASIS) is a project aimed at making neuroimaging data sets of the brain freely available to the scientific community. We use OASIS 2 of this dataset, which has longitudinal data for evaluation. This dataset consists of 335 T1-weighted sMRI scans collected from 135 subjects, including both AD subjects and healthy volunteers. We have a similar pre-processing step to ADNI. As a result, we also have images of size 224 × 224 × 224 with a resolution of $1.75mm \times 1.75mm \times 1.75mm$.

AIBL [13]. The Australian Imaging, Biomarker & Lifestyle flagship study of aging (AIBL) is a study to discover biomarkers, cognitive characteristics, and health and

Method	ADNI		OASIS		AIBL	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
3D ResNet50 [14]	82.72%	81.04%	66.15%	65.94%	72.99%	51.04%
3D ResNet101 [14]	85.19%	83.57%	69.23%	69.02%	76.43%	57.94%
3D ResNet152 [14]	87.65%	86.92%	70.77%	70.49%	77.59%	61.95%
3D DenseNet121 [18]	88.89%	87.98%	72.31%	72.68%	82.76%	73.26%
3D ViT [10]	80.24%	81.35%	67.69%	67.13%	73.56%	76.60%
MRNet [4]	87.96%	93.16%	70.77%	81.97%	75.86%	75.17%
MedicalNet [7]	88.89%	88.80%	73.85%	72.72%	82.76%	79.07%
M3T [20]	90.05%	88.78%	80.47%	81.67%	82.35%	80.26%
Single Image	90.67%	89.03%	80.23%	80.71%	82.78%	78.42%
w/ Prior Image (see Fig. 3)	91.42%	91.37%	81.38%	82.14%	83.94%	84.14%
LongFormer w/ VoxelMorph (ours)	92.70%	**93.75%**	81.46%	81.50%	**85.77%**	**84.47%**
LongFormer w/ Optical Flow (ours)	**93.43%**	93.30%	**82.35%**	**82.86%**	84.09%	84.24%

Table 1. Quantitative results of our LongFormer and baseline methods on classifying AD and NC subjects from three datasets. The top eight baselines are existing methods and the flowing four methods are variants of our Longformer for ablation study. The best results are in **bold** and the second best ones are colored in blue.

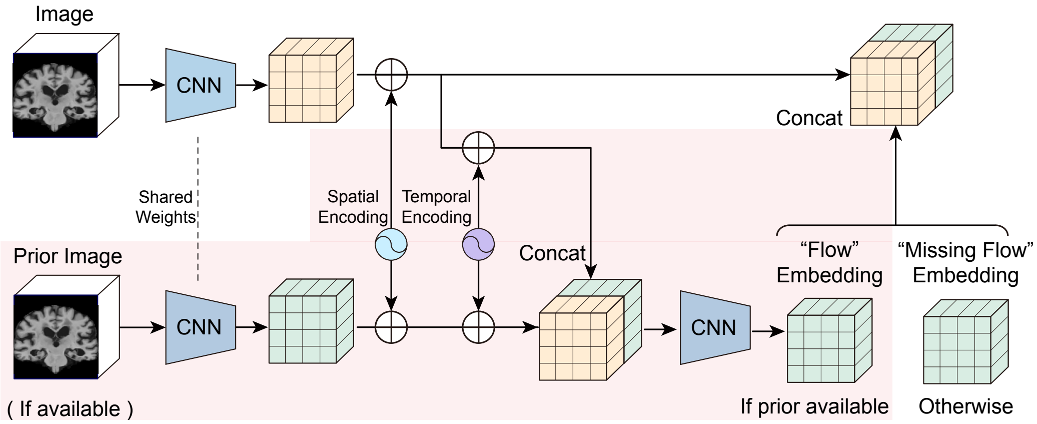

Figure 3. Ablation Study: Modifying the embedding module of our LongFormer to take a prior image instead of a pre-computed longitudinal flow. This architecture is similar to [3].

lifestyle factors that determine the subsequent development of symptomatic Alzheimer's Disease (AD). This dataset includes 997 T1-weighted structural MRI (sMRI) scans collected from 456 subjects with Alzheimer's disease (AD) and healthy volunteers. We separate training and evaluation sets just like ADNI, and use the raw data downloaded from the official website without any pre-processing, except for a simple alignment with centering the brain and normalizing the image intensity. Then, we resample the image volumes, resulting in images of size 224×224×224, with a resolution of $1.6mm \times 0.9mm \times 0.9mm$.

For all images, we normalize the image intensity to zero mean and unit variance. For all datasets, we *subject-wisely* divide them into 80% for training and 20% for testing.

4.2. Experimental Settings

We adopt a 3D DenseNet121 [18] as our CNN backbone to extract visual features from input pairs. The 3D CNN inputs 3D volumes or vector fields of size 224×224×224 and summarizes them into 3D representation features of size $7 \times 7 \times 7$ with 1024 channels. In the transformer part, we use $L = 6$ querying blocks and 125 learnable queries. The deformable cross-attention layer has a hidden dimension of 512 and 8 attention heads.

We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 100 epochs with a learning rate of 5e-5, and the batch size is 8. Our model is implemented using PyTorch-1.12 and is trained on four NVIDIA GeForce RTX 3090 GPUs. To evaluate the classification performance, we use the classification accuracy (ACC) and the area under receiver oper-

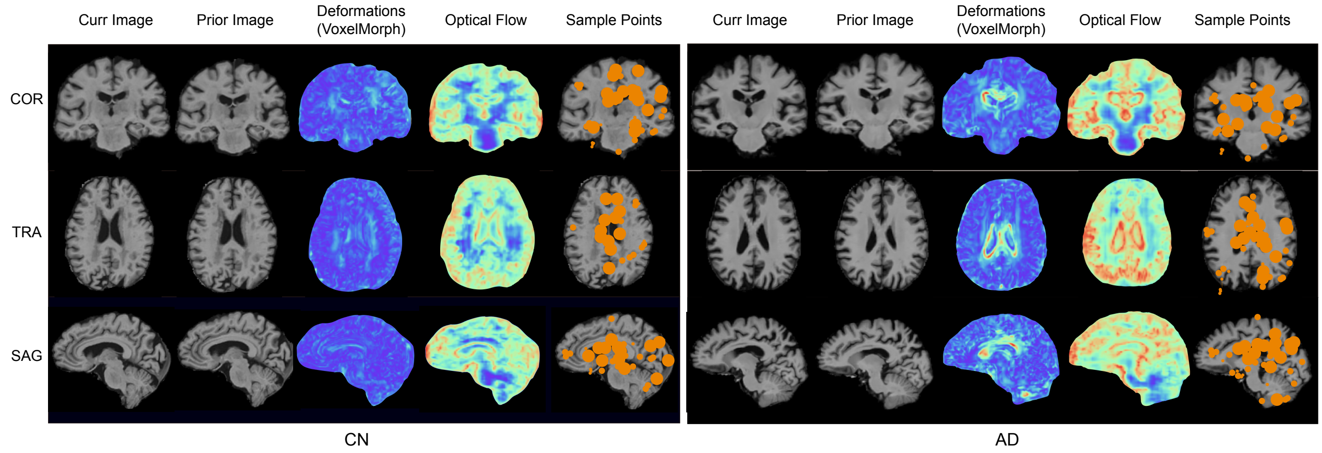

Figure 4. Sagittal (SAG), transverse (TRA), and coronal (COR) view visualization of longitudinal flows and deformable attention of our LongFormer, for AD and NC subjects sampled from the ADNI test set. The flows are visualized using their magnitudes. We use the jet colormap in that the red color is close to one (indicating high activated value) and the blue close to zero (indicating low activated value). The orange circles show the sample points with the highest propagated attention scores at multiple heads. A larger radius indicates a higher score. (Best viewed in color)

ating characteristic curve (AUC) as our evaluation metrics.

4.3. Comparison Results

We compare our LongFormer with conventional 3D classification methods based on 3D ResNet (50, 101, 152) [14], 3D DenseNet121 [17], and ViT [10], since these methods have been widely used for AD classification [12, 21, 22, 24, 30, 37, 39]. We implement the 3D ViT, which is composed of pure-transformer networks [10], and M3T [20] which is composed of 2D CNN, 3D CNN, and transformer networks. In the M3T model, the sequence in the transformer is applied to extracted 3D patch embedding with a size of $16 \times 16 \times 16$, and the projection dimension is 512. We also compare our model with MRNet [4] and MedicalNet [7]. The MRNet used in our experiment is based on 2D ResNet50 because it has better performance than AlexNet. MedicalNet is pre-trained on 23 medical databases.

Table 1 reports the quantitative results for comparing our method with baselines. Overall, our LongFormer performs best among these methods on all datasets. Even only considering a single image, our CNN-Transform hybrid design achieves comparable or better performance than the current SOTA method, M3T. Compared to 3D ViT, our method has a large performance gain, over 10% improvement on average, which indicates the effectiveness of using CNN as the backbone. We argue that a pure transformer needs a large amount of data for training, while our hybrid design alleviates this data demand greatly, which suits our datasets.

Figure 4 visualizes the longitudinal flows generated by VoxelMorph and optical flow, respectively, and the responses of our LongFormer on classifying AD and CN subjects. Since our LongFormer with VoxelMorph and optical flow produce similar responses, we only present the one using VoxelMorph. This response map visualizes those sample points in the deformable cross-attention layer, which are located at those important regions for recognizing AD and support our hypothesis in Fig. 1 at the beginning.

4.4. Ablation Study

(1) Longitudinal Flow. Firstly, we demonstrate the necessity of using the pre-computed longitudinal flow and the way of computing it. Therefore, we consider four cases: (1) using a single image, that is, whether we need longitudinal images for AD classification, or a cross-sectional setting is enough; (2) using the prior image directly, that is, whether the network can figure out the longitudinal changes from the image sequence directly; (3) using deformation fields computed by VoxelMorph, which is one way to compute the longitudinal flow; and (4) using optical flow to compute the flow, an alternative.

Specifically, for the second case, we modify our LongFormer network and implement a version of learning from image sequences directly, according to [3]. In this variant, after the spatial position encoding, we add a temporal encoding $\mathbf{T} \in \mathbb{R}^{T \times C_S}$, which takes into account the time information (see Figure 3). This modified embedding module outputs the concatenation of two features, which is an ‘aggregated’ representation of longitudinal flow features anchored on the current and prior image features.

As reported in Table 1, experiments with longitudinal settings outperform the one with a single image, which indicates longitudinal information helps in AD classification.

Backbone	ADNI		OASIS		AIBL	
	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
3D ResNet50	87.33 (+4.61)	85.41 (+4.37)	70.59 (+4.44)	73.33 (+7.39)	70.45 (+2.54)	72.62 (+21.58)
3D ResNet101	86.50 (+1.31)	86.23 (+2.66)	76.47 (+7.24)	71.43 (+2.41)	77.27 (+0.84)	79.40 (+21.46)
3D ResNet152	92.67 (+5.02)	92.64 (+5.72)	79.94 (+9.17)	80.71 (+10.22)	81.82 (+4.23)	80.40 (+18.45)
3D DenseNet121	**93.43** (+4.54)	**93.30** (+5.32)	**82.35** (+10.04)	**82.86** (+10.18)	**84.09** (+1.33)	**84.24** (+10.98)

Table 2. Classification performance comparison using different 3D CNN backbones in our LongFormer with optical flow. The numbers in red indicate the improved performance, compared to using the backbone network for classification directly, as reported in Table 1.

#Queries	ADNI		OASIS		AIBL	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
27	89.40%	88.72%	80.46%	81.99%	83.91%	79.22%
64	92.59%	91.58%	81.61%	81.67%	**84.48%**	80.26%
125	93.43%	**93.30%**	**82.35%**	**82.86%**	84.09%	**84.24%**
343	**93.82%**	92.64%	**82.35%**	80.71%	82.76%	82.09%

Table 3. Classification performance comparison using different numbers of queries in our LongFormer with optical flow. The best results are in **bold** and the second best ones are colored in blue.

Compared with directly working on image sequences, our model performs better when taking pre-computed longitudinal flows for learning. The two methods for computing the longitudinal flows, e.g., VoxelMorph and optical flow, perform equally well on our datasets.

(2) Image Embedding. Next, we experiment with the effect of our CNN backbone on AD classification. In this ablation study, we choose LongFormer with optical flow and compare four different backbone networks. As shown in Table 2, a deeper ResNet provides a better image embedding, while our choice, i.e., DenseNet121, performs best among these four backbones. This experiment demonstrates the backbone network plays an essential role in our LongFormer. Also, compared to directly using these backbones for classification as reported in Table 1, our LongFormer with transformer further improves the classification performance in all cases. This indicates the effectiveness of using a transformer to integrate longitudinal changes.

(3) Learnable Queries. Lastly, we would like to explore an optimal number of learnable queries for our LongFormer. As reported in Table 3, we test on four different numbers of queries, and the one having 125 queries performs the best for most cases, which is set as default for other experiments.

5. Conclusion and Discussion

In this paper, we have proposed an effective CNN-Transformer architecture, *LongFormer*, for Alzheimer's disease classification based on longitudinal sMRI volumes. LongFormer adopts attention mechanisms with learnable queries and deformable cross-attention to integrate both spatial and temporal information in an image scan and its longitudinal flow. Our proposed method provides a way to address the issues of missing data, the limited size of a dataset, and the subtle changes over time and subject differences in AD classification based on 3D longitudinal MRIs. The ablation studies demonstrate the effectiveness of our model design. Compared to multiple recent baselines, our model achieves the SOTA AD classification performance on three public datasets.

Limitations and Future Work. Currently, we only consider two scans of a subject, i.e., a current image and its prior one, to compute the longitudinal flow. To have a more accurate estimation of the flow, we could apply image regression [9, 15] on an image sequence. Further improving the quality of the estimated longitudinal flow would help improve the classification performance of our LongFormer. Also, our experiments focus on binary classification, which can be straightforwardly extended to multi-label classification, e.g., classifying AD, Mild Cognitive Impairment (MCI), and NC. Besides, our LongFormer only takes MRI scans for AD classification; however, other attributes, like age, gender, lab results, and other image modalities, like PET, fMRI, etc., are all beneficial for AD diagnosis. How to integrate them in a uniform framework for our task is an interesting research topic and left as our future work.

Acknowledgments

This work was supported by NSFC 62203303 and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 2, 4

[3] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. *arXiv preprint arXiv:2301.04558*, 2023. 6, 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. 6, 7

[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

[6] Dennis Chan, Nick C Fox, Rachael I Scahill, William R Crum, Jennifer L Whitwell, Guy Leschziner, Alex M Rossor, John M Stevens, Lisa Cipolotti, and Martin N Rossor. Patterns of temporal lobe atrophy in semantic dementia and alzheimer's disease. *Annals of neurology*, 49(4):433–442, 2001. 1

[7] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 6, 7

[8] Ruoxuan Cui, Manhua Liu, Alzheimer's Disease Neuroimaging Initiative, et al. Rnn-based longitudinal analysis for diagnosis of alzheimer's disease. *Computerized Medical Imaging and Graphics*, 73:1–10, 2019. 2, 3

[9] Zhipeng Ding, Greg Fleishman, Xiao Yang, Paul Thompson, Roland Kwitt, Marc Niethammer, Alzheimer's Disease Neuroimaging Initiative, et al. Fast predictive simple geodesic regression. *Medical image analysis*, 56:193–209, 2019. 8

[10] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6, 7

[11] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 4

[12] Amir Ebrahimi, Suhuai Luo, and Raymond Chiong. Introducing transfer learning 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. 7

[13] Kathryn A Ellis, Ashley I Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T Lautenschlager, Nat Lenzo, Ralph N Martins, Paul Maruff, et al. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International psychogeriatrics*, 21(4):672–687, 2009. 1, 2, 5

[14] 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. 4, 6, 7

[15] Yi Hong, Polina Golland, and Miaomiao Zhang. Fast geodesic regression for population-based image analysis. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 317–325. Springer, 2017. 8

[16] Zhentao Hu, Zheng Wang, Yong Jin, and Wei Hou. Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction. *Computer Methods and Programs in Biomedicine*, 229:107291, 2023. 1, 2, 3

[17] 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. 4, 7

[18] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 6

[19] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008. 1, 2, 5

[20] Jinseong Jang and Dosik Hwang. M3t: Three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20718–20729, 2022. 1, 2, 6, 7

[21] Hiroki Karasawa, Chien-Liang Liu, and Hayato Ohwada. Deep 3d convolutional neural network architectures for alzheimer's disease diagnosis. In *Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19-21, 2018, Proceedings, Part I 10*, pages 287–296. Springer, 2018. 7

[22] 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. 7

[23] Chao Li, Yue Cui, Na Luo, Yong Liu, Pierrick Bourgeat, Jurgen Fripp, and Tianzi Jiang. Trans-resnet: Integrating transformers and cnns for alzheimer's disease classification.

In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022. 1, 2

[24] Qi Li and Mary Qu Yang. Comparison of machine learning approaches for enhancing alzheimer's disease classification. *PeerJ*, 9:e10549, 2021. 7

[25] Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):880–893, 2018. 2

[26] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007. 1, 2, 5

[27] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 4

[28] Anqi Qiu, Liyuan Xu, Chaoqiang Liu, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting diagnosis 4 years prior to alzheimer's disease incident. *NeuroImage: Clinical*, 34:102993, 2022. 1

[29] 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. 2

[30] Juan Ruiz, Mufti Mahmud, Md Modasshir, M Shamim Kaiser, and for the Alzheimer's Disease Neuroimaging Initiative. 3d densenet ensemble in 4-way classification of alzheimer's disease. In *Brain Informatics: 13th International Conference, BI 2020, Padua, Italy, September 19, 2020, Proceedings 13*, pages 85–96. Springer, 2020. 7

[31] Krishnakant Saboo, Anirudh Choudhary, Yurui Cao, Gregory Worrell, David Jones, and Ravishankar Iyer. Reinforcement learning based disease progression model for alzheimer's disease. *Advances in Neural Information Processing Systems*, 34:20903–20915, 2021. 1

[32] Paul M Thompson, Kiralee M Hayashi, Greig De Zubicaray, Andrew L Janke, Stephen E Rose, James Semple, David Herman, Michael S Hong, Stephanie S Dittmer, David M Doddrell, et al. Dynamics of gray matter loss in alzheimer's disease. *Journal of neuroscience*, 23(3):994–1005, 2003. 1

[33] 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. 2

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5

[35] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022. 2

[36] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022. 2, 4

[37] 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. 7

[38] Gengshan Yang and Deva Ramanan. Upgrading optical flow to 3d scene flow through optical expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1334–1343, 2020. 2, 4

[39] 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. 7

[40] Shengjie Zhang, Xiang Chen, Bohan Ren, Haibo Yang, Ziqi Yu, Xiao-Yong Zhang, and Yuan Zhou. 3d global fourier network for alzheimer's disease diagnosis using structural mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 34–43. Springer, 2022. 1, 2

[41] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[42] Wenyong Zhu, Liang Sun, Jiashuang Huang, Liangxiu Han, and Daoqiang Zhang. Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri. *IEEE Transactions on Medical Imaging*, 40(9):2354–2366, 2021. 1, 2

[43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2