2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782744

Integrating Vision Transformer with UNet++ for Hippocampus Segmentation in Alzheimer's Disease

Thony Yan Liang, Christian Freytes, Xueting Cui, Bipul Simkhada, Marcos Bosques-Perez,
Mercedes Cabrerizo, Armando Barreto, Malek Adjouadi
Florida International University College of Engineering & Computing
10555 W Flagler St, Miami, FL 33174
{tyan001, xcui005, bsimk001, cfrey001, mbosq005, cabreriz, barretoa, adjouadi}@fiu.edu

***Abstract*—The hippocampus is a disease-prone area of the brain that can be used as an important biomarker for neurodegenerative diseases like Alzheimer's. In recent years, deep neural networks have been applied to segment the hippocampus. However, accurately segmenting the hippocampus using magnetic resonance imaging (MRI) remains a challenging task. To explore a more effective segmentation strategy, this study proposes a new model by integrating the Vision Transformer (ViT) architecture with the UNet++ architecture, which is validated by using manual tracing of the hippocampus performed by clinical experts. The proposed ViT-based model achieved a dice score of 0.885, surpassing similar models by 2.82% in the Dice coefficient score.**

***Index Terms*—Manual tracing, UNet++, Hippocampus, Segmentation, Dice Score**

I. Introduction

The hippocampus is a region of interest (ROI) for various research studies, including memory function analysis and for observing and predicting neurological and neurodegenerative disorders [1]–[3]. Tracking hippocampal atrophy through accurate volumetric calculations could help classify and predict AD via deep learning models. In medical imaging, the standard practice for disease-prone regions, such as the hippocampus, is manual tracing performed by clinical experts [4]. This manual process is time-consuming and ultimately highly subjective, making it error-prone. These errors can be due to the resolution of the Magnetic Resonance Imaging (MRI) scans and the difficulty in delineating a given region in different scans while considering all the possible variations in its structural form. This variation is expressed in a study by Boccardi where when four expert tracers approved of each other's structural accuracy, their volumetric measurements still revealed a mean difference of 9% with a variation of 7% from the estimated actual volumes [5]. Supplementing deep learning models with manual tracings such as these could potentially increase the preciseness of volumetric measurements.

MRI hippocampal segmentation is an essential procedure for prognosis, providing volumetric and structural data [6], [7]. With this information, a medical provider can diagnose various diseases, including Alzheimer's Disease and Dementia [8]. Contemporary hippocampal segmentation often includes three main methods: manual, semi-automated, and fully automated.

The U-net architecture is a leading deep-learning model for segmentation tasks [9] and is frequently implemented for medical image segmentation. Various iterations of the U-net emerged to improve its architecture and performance. Among them is the UNet++ that improves upon the standard U-net architecture by re-designing the skip path connections, enabling information flow across multiple layers [10]. Although these models are exceptional, the foundation of the models is convolutional neural networks (CNN) that typically exhibit limitations in information retention due to their depth and reliance on numerous convolutional layers known as the vanishing gradient problem [11]. In some studies, a self-attention mechanism was implemented into CNN-based architecture to address the problem [12], [13].

Transformers have become increasingly popular among researchers for visual tasks. Recently, a new type of architecture has emerged that demonstrates a comparable performance with state-of-the-art techniques for image tasks called Vision Transformer (ViT) [14]. This paper presents an architecture implementing ViT architecture into the UNet++ model and explores the viability of using ViT architecture within U-Net base models for Hippocampus image segmentation. This combined architecture will be called TransUNet++.

II. Related Work

There have been various studies using ViT with U-Net based architecture. Jieneng *et al*. first introduced TransUnet for semantic segmentation and proposed implementing the ViT architecture in the U-Net architecture bottleneck section, showing that, on average, it outperformed other base models by about 6.36% in dice scores, such as U-Net and attention-based U-Net [15]. Hatamizadeh used the final output from ViT model with no encoding layers and then used intermediate outputs generated within the ViT architecture. These intermediate outputs were then used as the data for the skip connection in the decoding section of their model [16]. This technique mimicked the behavior of the U-Net architecture by facilitating information flow across the model.

Studies have also been conducted on the impact of small sample data and deep networks. Xu *et al*. created FedSM to solve the generalization gap from insufficient data due to strict medical imaging data sharing rules [17]. Amin *et al*. implemented their own ViT U-net architecture to address *Catastrophic Forgetting* [13] in CNNs and using transfer

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 15,2025 at 13:36:12 UTC from IEEE Xplore. Restrictions apply.

learning for their ViT architecture to prove that Transformer can be used with U-Net and not jeopardize its effectiveness and prevent the model from not retaining information as it goes through the architecture [18].

III. Model

A. UNet++

UNet++ architecture re-designed the skip pathway from the original U-Net architecture. Instead of collecting features directly from the encoder at the same pyramid level, UNet++ undergoes different convolutional operations with other feature maps from lower levels through up-sampling operations [10]. With this approach, the dense convolutional operation yields improved feature maps once it reaches the decoders. Through this method, the architecture's optimizer can outperform the standard skip-connection approach from the U-Net architecture.

The UNet++ skip pathway is formulated as follows: let $x^{i,j}$ be the output node of $X^{i,j}$ where i is the index for the down-sampling layer along the encoder path and j index the convolution layers along the skip-connection pathways. The way to compute the $x^{i,j}$ for the stack of feature maps will be as follows [10]:

$$x^{i,j} = \begin{cases} \mathcal{C}\left(x^{i-1,j}\right), & j = 0 \\ \mathcal{C}\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, \mathcal{U}\left(x^{i+1,j-1}\right)\right]\right), & j > 0 \end{cases} \quad (1)$$

The function $C(*)$ is a convolutional block operation followed by an activation function. $U(*)$ is an up-sampling operation to then concatenate with layers above denoted by $\left[*\right]$.

B. Vision Transformer

The ViT architecture is implemented after the encoding pathway of the UNet++ architecture following the methodology proposed by [19]. It begins by tokenizing the last encoding output, denoted as $\mathbf{x}$, into a sequence of flattened 2D patches, each with a $P \times P$ dimension. The total number of patches is $N = H \times W/P^2$. Where H and W represent the height and width of the feature maps. This sequence of patches will serve as the input for the ViT model. Next, patch embedding is performed, mapping each patch x_p into a D-dimensional embedding space using trainable linear projections. To retain information, positional embeddings are added to the patch embeddings, following the approach introduced by Chen [15]. The resulting equation for this process is as follows:"

$$\mathbf{z}_0 = \left[\mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos} \quad (2)$$

The transformer encoder consists of multiple layers denoted as L, incorporating multi-headed self-attention (MSA), Multilayer Perceptrons (MLP), and Layernorm (LN). The output of the ℓ layer is expressed as follows:

$$\mathbf{z}'_\ell = \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{z}_{\ell-1}\right)\right) + \mathbf{z}_{\ell-1} \quad (3)$$

$$\mathbf{z}_\ell = \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{z}'_\ell\right)\right) + \mathbf{z}'_\ell \quad (4)$$

C. TransUNet++

The TransUNet++ architecture combines the strengths of two models: UNet++ and ViT. UNet++ has effectively captured multi-scale features through its re-designed skip connections [10], while ViT captures long-range dependencies and global context. This integration aims to create a model that can retain information as it flows through the network, leading to better image segmentation. One essential part of the TransUNet++ model is the output of the ViT architecture, which leads to the bottleneck section of the architecture. To revert to the decoding layers' spatial resolution, the output embedding features (N, D) will be reshaped to $(D, H/16, W/16)$. Afterwards, a 1x1 convolution kernel convolution operation is implemented to create feature maps for the decoding layers. The architecture can be seen in Fig. 1.

IV. Method

The first objective is pre-processing data by extracting the left and right hippocampus. Random K-fold cross-validation is then utilized to generate *K* different datasets that have been randomly split into train, testing, and validation sets. These training and validation datasets will be used to train the TransUNet++ model, and the testing dataset will be utilized to evaluate the model's performance.

A. Dataset

The *Harmonized Hippocampal Protocol* (HarP) dataset [20] consist of 135 subjects categorized into three clinical groups: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). Tab. I presents the number of subjects for each class of this dataset. From the 135 scans, 68 are 1.5T, and 67 are 3T volumetric structural scans from different subjects. Five qualified HarP tracers were tasked to manually segment the hippocampus from these 135 scans [21]. This resulted in absolute interrater intraclass correlation coefficients of 0.953 and 0.975 for the left and right hippocampus, respectively [5].

TABLE I: HarP dataset class distribution

Dataset Name	CN	MCI	AD	Total
HarP	44	46	45	135

Hippocampal volume is calculated using the corresponding hippocampus segmentation made by the clinical experts to compare the volumetric data for each clinical group (CN, MCI, AD). As each pixel represents a 1mm cubic voxel size, calculating the total number of pixels per hippocampus will give the volumetric result of the left and right hippocampus. Fig. 2 illustrates a trend in which the average hippocampus volume declines as the disease progresses through its different stages.

B. Pre-Processing

The hippocampus is extracted from each subject using the labeled data. This is done by identifying the center point between the left and right hippocampus and then cropping the

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 15,2025 at 13:36:12 UTC from IEEE Xplore. Restrictions apply.

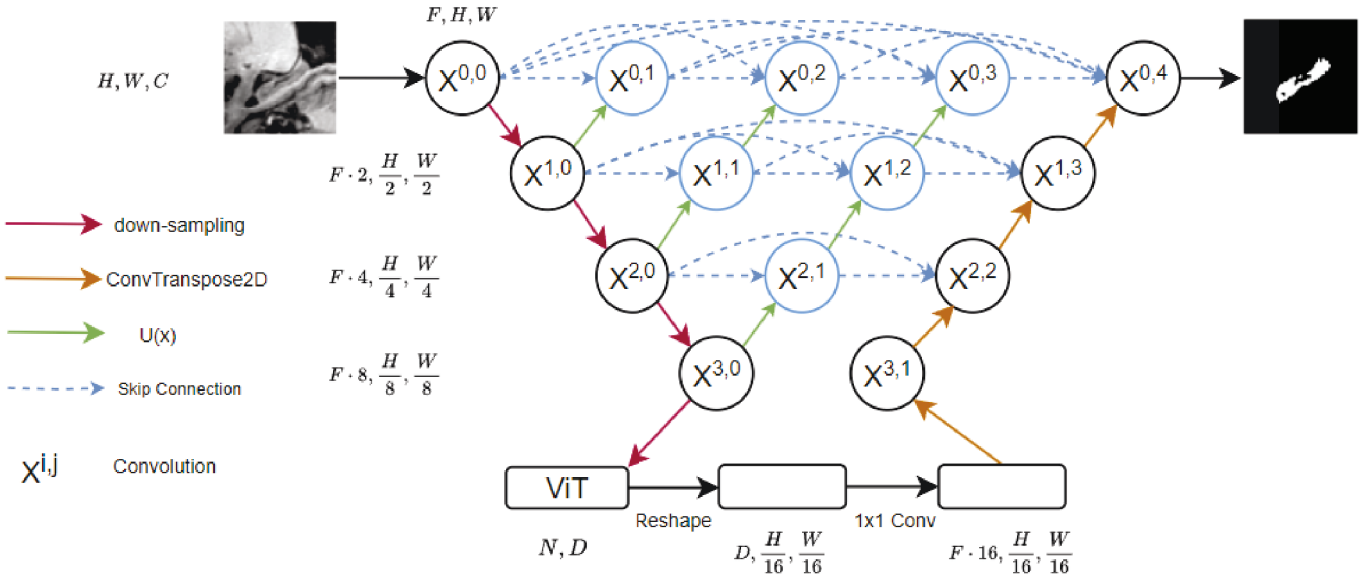


Fig. 1: TransUNet++ architecture overview.

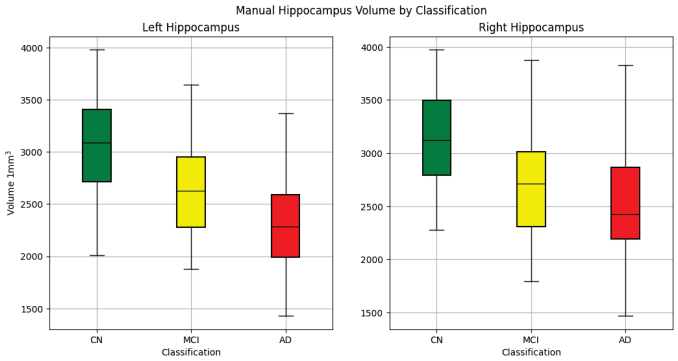
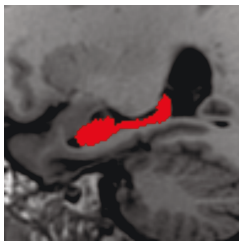


Fig. 2: Manual hippocampus volumes for each classification

image to the size of $96 \times 64 \times 64$ around the calculated center point. The cropped dataset is then normalized per subject. The output of this pre-processing is depicted in Fig. 3. This study focuses only on segmenting the hippocampus for the sagittal axis of the brain scan.



(a) MRI Slice (b) Cropped MRI Slice

Fig. 3: (a) a MRI slice of a patient retrieved from the HarP dataset. (b) The cropped data was retrieved from the same patient. The red pixels are the label data for the hippocampus. Viewed using Mango.

C. Cross Validation

K-fold random cross-validation is a strategic approach to assess the proposed model's stability and robustness that is applied to the dataset. Resulting in five randomly split datasets with a ratio split of 60, 20, and 20 for training, testing, and validating, respectively. This is done for the following reasons:(1) To explore the stability of the model with randomized distribution of data classes. (2) To assess if the model would generalize well across all five randomized datasets. (3) to evaluate if the hyperparameters are viable across all dataset splits. In this experiment, 5-fold random cross-validation is performed on the dataset.

D. Loss Function

The loss function used for this experiment is a fusion of the Binary Cross-Entropy (BCE) loss and the dice loss, which will be referred to as BCE Dice Loss [12]. The BCE Loss is a standard loss function for binary classification, and it measures the dissimilarity between predicted probabilities and the ground truth. The dice loss addresses class imbalance regarding pixel-wise classification. Due to the fact that the label mask is mostly black pixels, the model will be biased toward that class. The dice loss evaluates overlaps between the predicted and ground truth, rewarding high overlaps and penalizing low overlaps. The integration of both loss functions is as follows:

$$L_{BCE}(y, \widehat{y}) = -(y \log(\widehat{y}) + (1 - y) \log(1 - \widehat{y})) \tag{5}$$

$$L_{DL}(y, \widehat{y}) = 1 - \frac{2y\widehat{y} + 1}{y + \widehat{y} + 1} \tag{6}$$

$$L_{BCEDICE}(y, \widehat{y}) = L_{BCE}(y, \widehat{y}) + L_{DL}(y, \widehat{y}) \tag{7}$$

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 15,2025 at 13:36:12 UTC from IEEE Xplore. Restrictions apply.

E. Metric

This experiment employs a key performance metric, the Dice coefficient. The significance of this metric lies in its ability to provide comprehensive insights into various aspects of our model's performance, thereby enabling a thorough evaluation. The Dice Coefficient measures the overlap between two samples. The equation is as follows:

$$\text{Dice Coefficient} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{8}$$

V. Training

A. Model's hyperparamters

For this experiment, the convolutional blocks are designed with a sequence of operations: a bias-free convolution, followed by a batch normalization and the use of the ReLU activation function. This process is iterated twice to form a convolutional block. Transpose Convolution is used to facilitate up-sampling for the decoding section of the model. For the ViT architecture, the hyperparameter is based on the ViT-Base from [14], where there are 12 Encoding layers with a Hidden size of D 768, an MLP size of 3072, and a Heads size of 12. The model is trained using the NAdam [22] optimizer $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a learning rate value of $1e^{-}3$ with an early stop mechanism to prevent overfitting to the training dataset.

B. Data Augmentation

Data augmentation improves model performance by creating dataset variation, improving the model's robustness and generalization ability. For this experiment, a random affine transformation was applied to an MRI slice that scales the image between 0.85 and 1.15, rotates it by 5 degrees, and can randomly flip horizontally.

VI. Results

After training the model, the testing dataset is used to see the model's performance. Since 5-fold random cross-validation is utilized, five distinct models are generated, each with their unique dice score value. All five were tabulated, and the mean and standard deviation were calculated. The overall dice score results can be observed in Tab. II, alongside results from similar models for comparison.

TABLE II: Dice score from models on the HarP dataset

Architecture	Dice $\pm\sigma$
nnUnet [18]	$85.72 \pm (0.77)$
TransUNet [18]	$85.74 \pm (0.99)$
UNet++	$86.59 \pm (0.016)$
TransUNet++	**88.56 $\pm$ (0.012)**

Different subjects were observed to thoroughly assess the model's performance on the testing dataset. The model demonstrates a strong dice score when predicting the hippocampus volume in the central region of the hippocampus. However, it encounters challenges when the boundaries between the hippocampus and its surrounding tissues become ambiguous. In Fig. 4, a side-by-side comparison of the manual tracing (represented as the True segmentation) and the model's predicted segmentation can be seen. For this MRI scan, the dice score was 0.95, but the dice scores decreased at the borderline of the hippocampus.

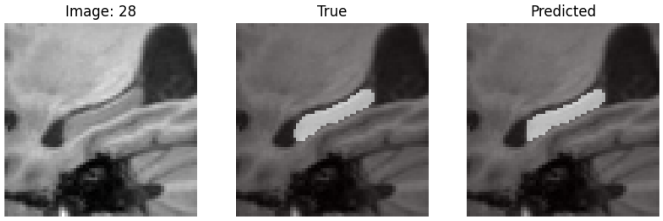


Fig. 4: Example output from the TransUNet++ model. The input image is shown on the left, the manual tracing is in the middle, and the predicted segmentation is on the right.

Further analysis was done using the top-performing TransUNet++ model to determine its accuracy in segmenting the hippocampus for certain cases. This was done using TransUNet++ to generate predictions on the test dataset. Tab. III displays the class distribution within the test dataset, along with the mean and standard deviation of the dice score.

TABLE III: HarP test dataset class sizes and mean dice scores

Classification	Quantity	Dice Score
CN	10	0.839 ± 0.146
MCI	6	0.890 ± 0.013
AD	11	0.884 ± 0.019

CN subjects had a larger standard deviation than the other clinical groups in the dice score. In one specific case, when observing the hippocampus frame by frame, it was apparent that it was not well segmented in some instances, as shown in Fig. 5. This discrepancy is what led to a lower dice score and higher standard deviation. In Tab. IV, the volumetric data of the hippocampus from the HarP dataset is compared with the prediction that is generated by the TransUNet++ model. The model predictions result in a higher volumetric value than the HarP segmentation across all clinical groups, which suggests that the model is more inclusive in predicting the extent of the hippocampus area, thus leading to a higher volume estimation of the hippocampus.

VII. Conclusion

This study proposes an automated approach to segmenting the hippocampus to overcome the tediousness of manual tracing and the inordinate amount of time it takes to trace the hippocampus. It should be noted that even with manual tracers, the difference between their tracing can vary up to 9% with a standard deviation of 7% [5] [20], which only confirms that segmentation of brain regions without any visible boundaries in MRI scans is a challenging task. The objective was to leverage these manual tracings to train a deep-learning model capable of mitigating the tracing variations among clinical experts. This is demonstrated in the architecture's performance when there was a small variation in the dice score when

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 15,2025 at 13:36:12 UTC from IEEE Xplore. Restrictions apply.

	HarP Segmentation						TransUNet++ Prediction					
Classification	CN		MCI		AD		CN		MCI		AD	
Hemisphere	LH	RH	LH	RH	LH	RH	LH	RH	LH	RH	LH	RH
Mean	3108	3185	2645	2732	2339	2486	3183	3255	2739	2821	2402	2542
std	532.00	497.82	463.65	479.75	497.32	548.46	505.43	467.66	471.77	480.68	500.05	548.46

TABLE IV: Hippocampal volumes of HarP dataset and TransUNet++ prediction

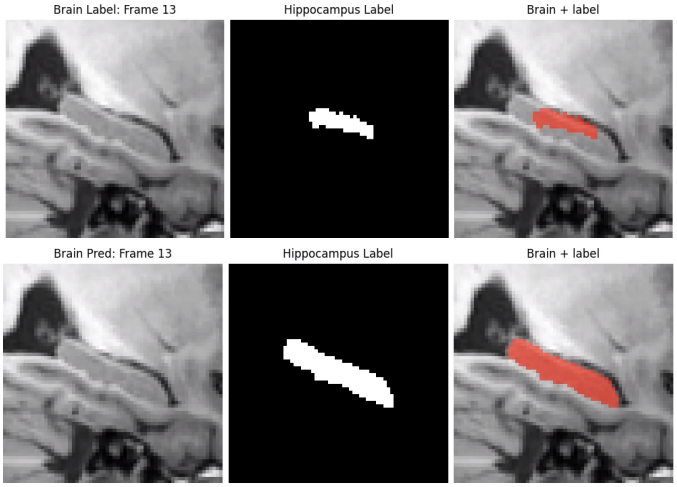


Fig. 5: Top three images are the input image, the true label mask, and the true label mask superimposed on the input image. The bottom three images are the input image, the predicted label mask, and the predicted label mask superimposed on the input image.

random k-fold cross-validation was implemented. However, a more extensive dataset will be needed to see the model's full potential. Despite employing data augmentation, the model's performance is still constrained by the limited dataset size.

This study explores the integration of the ViT within the UNet++ architecture and evaluates its performance for image segmentation tasks. By combining UNet++'s ability to capture multi-scale features with its redesigned skip connections and the ViT architecture excelling in long-range dependency and global context, the culminated architecture led to more accurate segmentation in the hippocampus. The results achieved by this architecture are encouraging, with a dice score of 0.885 ± 0.012. This score showed an improvement of 2.82% compared to other similar models.

ACKNOWLEDGMENT

This research is supported by the National Science Foundation under grants: CNS-1920182, CNS-2018611, CNS-1551221, and with the National Institutes of Health, National Institute on Aging (NIA) through the P30AG066506-01 with the 1Florida Alzheimer's Disease Research Center (ADRC). We also acknowledge the support of the Ware Foundation.

REFERENCES

[1] J. C. R. Jr, P. RC, X. Y, O. PC, S. GE, and *et al*, "Rates of hippocampal atrophy correlate with change in clinical status in aging and ad," *Neurology*, vol. 55, no. 4, 2000.

[2] P. Kathryn, K. Richard, S. Betg, M. Nicola, W. Dorothy, and *et al*, "Processing speed in normal aging: effects of white matter hyperintensities and hippocampal volume loss," *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn.*, vol. 21, no. 2, 2014.

[3] R. Aaron, B. Adam, M. Jordan, S. Jason, and S. Yaakov, "Hippocampal atrophy relates to fluid intelligence decline in the elderly.," *J Int Neuropsychol Soc.*, vol. 17, no. 1, 2011.

[4] M. Boccardi, M. Bocchetta, R. Ganzola, N. Robitaille, A. Redolfi, S. Duchesne, and *et al*, "Operationalizing protocol differences for eadc-adni manual hippocampal segmentation," *Alzheimer's & Dementia*, vol. 11, no. 2, pp. 184–194, 2015.

[5] M. Boccardi, M. Bocchetta, F. C. Morency, D. L. Collins, M. Nishikawa, R. Ganzola, and *et al*, "Training labels for hippocampal segmentation based on the eadc-adni harmonized hippocampal protocol," *Alzheimer's & Dementia*, vol. 11, no. 2, pp. 175–183, 2015.

[6] J. C. Jr, K. DS, J. WJ, S. LM, A. PS, and *et al*, "Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade.," *Lancet Neurol*, vol. 9, no. 1, 2010.

[7] J. Bruno, L. Andrew, L. Bo, K. Elyse, Z. Yanwei, and *et al*, "A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort.," *Neuroimage*, vol. 63, no. 3, 2012.

[8] V. Dill, A. R. Franco, and M. S. Pinho, "Automated methods for hippocampus segmentation: the evolution and a review of the state of the art," *Neuroinform*, vol. 13, 133-150, 2014.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[10] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *CoRR*, vol. abs/1807.10165, 2018.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[12] X. Cui, T. Y. Liang, M. Aghili, M. Adeyosoye, R. E. C. Cid, D. Lowenstein, and *et al*, "Unet++ with attention mechanism for hippocampus segmentation," in *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1530–1534, 2022.

[13] C. Gonzalez, G. Sakas, and A. Mukhopadhyay, "What is wrong with continual learning in medical image segmentation?," *CoRR*, vol. abs/2010.11008, 2020.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, and *et al*, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, and *et al*, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021.

[16] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, and *et al*, "Unetr: Transformers for 3d medical image segmentation," 2021.

[17] A. Xu, W. Li, P. Guo, D. Yang, H. Roth, A. Hatamizadeh, and *et al*, "Closing the generalization gap of cross-silo federated medical image segmentation," 2023.

[18] A. Ranem, C. González, and A. Mukhopadhyay, "Continual hippocampus segmentation with transformers," 2022.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and *et al*, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[20] G. B. Frisoni, C. R. Jack, M. Bocchetta, C. Bauer, K. S. Frederiksen, and Y. L. and*et al*, "The eadc-adni harmonized protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity," *Alzheimer's & Dementia*, vol. 11, no. 2, pp. 111–125, 2015.

[21] M. Bocchetta, M. Boccardi, R. Ganzola, L. Apostolova, G. Preboske, and *et al*, "Harmonized benchmark labels of the hippocampus on magnetic resonance: the eadc-adni project," *Alzheimer's & Dementia*, vol. 11, no. 2, pp. 151–160, 2015.

[22] T. Dozat, "Incorporating nesteroc momentum into adam," ICLR, 2016.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on July 15,2025 at 13:36:12 UTC from IEEE Xplore. Restrictions apply.