BMC Neurology

RESEARCH

Open Access

Automatic segmentation of white matter lesions on multi-parametric MRI: convolutional neural network versus vision transformer

Yun-Ting Chen[1†], Yan-Cheng Huang[2†], Hsiu-Ling Chen[1], Hsin-Chih Lo[2], Pei-Chin Chen[1], Chiun-Chieh Yu[1], Yi-Chin Tu[2], Tyng-Luh Liu[2] and Wei-Che Lin[3*]

Abstract

Background and purpose White matter hyperintensities in brain MRI are key indicators of various neurological conditions, and their accurate segmentation is essential for assessing disease progression. This study aims to evaluate the performance of a 3D convolutional neural network and a 3D Transformer-based model for white matter hyperintensities segmentation, focusing on their efficacy with limited datasets and similar computational resources.

Materials and methods We implemented a convolution-based model (3D ResNet-50 U-Net with spatial and channel squeeze & excitation) and a Transformer-based model (3D Swin Transformer with a convolutional stem). The models were evaluated on two clinical datasets from Kaohsiung Chang Gung Memorial Hospital and National Center for High-Performance Computing. Four metrics were used for evaluation: Dice similarity coefficient, lesion segmentation, lesion F1-Score, and lesion sensitivity.

Results The Transformer-based model, with appropriate adjustments, outperformed the well-established convolution-based model in foreground Dice similarity coefficient, lesion F1-Score, and sensitivity, demonstrating robust segmentation accuracy. DRLoc enhanced the Transformer's performance, achieving comparable results on internal and benchmark datasets despite limited data availability.

Conclusion With comparable computational overhead, a Transformer-based model can surpass a well-established convolution-based model in white matter hyperintensities segmentation on small datasets by capturing global context effectively, making them suitable for clinical applications where computational resources are constrained.

Keywords White matter hyperintensities, Segmentation, Brain MRI, Convolutional neural network, Vision transformer

[†]Yun-Ting Chen and Yan-Cheng Huang share the role of first author.

*Correspondence:
Wei-Che Lin
alex@cgmh.org.tw
[1]Department of Diagnostic Radiology, Kaohsiung Chang Gung Memorial Hospital, Chang Gung University College of Medicine, No. 123 Ta-Pei Road, Niao-Sung Dist, Kaohsiung City 83305, Taiwan
[2]Taiwan AI Labs, 6F., No. 70, Sec. 1, Chengde Rd., Datong Dist, 103622 Taipei City, Taiwan
[3]Department of Diagnostic Radiology, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, School of Medicine, College of Medicine, National Sun Yat-Sen University, No. 123 Ta-Pei Road, Niao-Sung Dist, Kaohsiung 83305, Taiwan

Introduction

White matter hyperintensities (WMHs) are abnormal clusters of T2-weighted hyperintense signals in the cerebral white matter associated with various pathologies and geriatric conditions. They serve as important predictors for diseases such as stroke, cognitive decline, dementia, and mortality [1–3]. The number, volume, shape, signal intensity, and spatial distributions of the WMHs vary, reflecting the unique characteristics of each condition [4]. Comprehensive assessment of WMHs can provide valuable insights into etiology, prognosis, disease progression, and treatment effectiveness [1].

Convolutional neural networks (CNNs) are among the most successful deep learning models in medical imaging tasks, excelling in mass detection and segmentation across domains such as mammography [5], brain tumors [6], prostate cancer [7], uterine fibroids [8], and WMH segmentation [1].

While many studies have focused on CNN-based medical image analysis [1, 9, 10], The majority of 3D medical image analysis still relies on 2D CNN models. However, 3D images with spatial information are more desirable for medical segmentation [9, 11]. Unlike 2D CNN, 3D CNNs work with four-dimensional input and output feature spaces (channel, depth, height, and width) and can extract inter-slice information from adjacent frames, which is crucial when regions of interest span multiple frames of 3D volumetric data. 3D CNN optimize vector multiplication for faster computation, though they involve more parameters per convolution kernel. However, 3D CNN models also have limitations. The complexity of 3D data requires more storage and larger datasets for accurate segmentation [9, 11, 12]. They also need task-specific labeled data, which is often time-consuming and inevitably generates label noise. Although CNNs perform well on small to medium-sized datasets, they struggle with capturing global and long-range semantic information [13]. To address this, self-attention mechanisms [14–16] have been introduced to help CNNs retrieve global information.

Recently, Transformer-based networks, which rely entirely on self-attention for input-output representation, have been explored for medical image tasks like brain tumor and organ segmentation [17–19]. Both CNNs and Transformers have strengths and weaknesses. Hybrid models, combining CNN and Transformer architectures with self-attention mechanisms (e.g., nnFormer [17], TransUNet [20]), have shown improved global context modeling without sacrificing the ability to capture low-level details. However, Transformer can introduce significant computational overhead and require large-scale datasets, which are often difficult to obtain in medical applications [21, 22].

Our study aimed to explore modern convolutional and Transformer architectures, and investigate how to train a Transformer model to match or surpass a well-established CNN model for WMH segmentation on small datasets, using similar computational resources. The UNet architecture, known for its flexibility, modular design, contextual information incorporation, and fast training speed, was chosen as the foundation [23–25]. We developed two models: a CNN-based model using 3D ResNet50 U-Net [26] with spatial and channel squeeze & excitation (scSE) [14] and a Transformer-based model based on 3D Swin Transformer [27] with a modified convolutional stem and upsampling/downsampling blocks. Both models were adjusted to achieve similar complexity.

This article outlines our step-by-step process of training a Transformer model on small datasets to achieve comparable or superior performance to mature CNN models for WMH segmentation. Our models were validated using the WMH challenge dataset from MICCAI 2017 [28], providing an accurate, robust tool for WMH segmentation on fluid-attenuated inversion recovery (FLAIR) images to assist clinicians and researchers in their work.

Materials and methods

Subjects

In this study, we retrospectively reviewed adult outpatients who underwent brain MRI at Kaohsiung Chang Gung Memorial Hospital (KCGMH) between January 2010 and December 2018. A total of 121 cases were collected from the Chang Gung Research Database. Of these, 62 patients (51%) were female, and 59 (49%) were male, with an average age of 60 years (range: 30 to 80 years). All patients were Asian. The study protocol was approved by the Institutional Review Board of Chang Gung Memorial Hospital (IRB No. 202002026A3, approved on 19 January 2021; IRB NO. 201900483B0, approved on 10 April 2019). All patient data were anonymized and de-identified.

Exclusion criteria included patients with.

1. Intracranial hemorrhage.
2. Intracranial space occupying lesions.
3. A history of craniotomy, craniectomy, or intracranial neurosurgery.
4. Reports mentioning transarterial embolization (TAE), seizure, epilepsy, arteriovenous malformation (AVM), arteriovenous fistula (AVF), intoxication, tumor, metastasis, cancer, multiple sclerosis (MS), radiotherapy, Parkison's disease, moyamoya, tuberous sclerosis, trauma, hypoxic encephalopathy, necrosis, and hydrocephalus.
5. Patients with poor FLAIR image quality.

Additionally, the second clinical dataset included 505 cases obtained from the National Center for High-Performance Computing (NCHC). All patients in the TMUH dataset were of Asian descent.

Image dataset

Three clinical datasets and one research dataset were used to explore and validate the characteristics of CNN-based and Transformer-based models. The first clinical dataset was acquired from KCGMH, and the second from the NCHC. FLAIR images from the NCHC dataset were collected from Taipei Medical University Hospital (TMUH) to assess small vessel disease. The third dataset is a mixed collection of KCGMH and NCHC data. The research dataset was acquired from the MICCAI WMH Challenge [28].

MRI acquisition

All brain MRIs from KCGMH were reviewed for high-quality FLAIR images. Scans were acquired from three vendors (GE, Philips, Siemens) and five different scanners with 1.5T or 3T field strengths and different average resolutions, following clinical FLAIR protocols. The images had an axial thickness of 2 to 5 mm. More details, including voxel sizes, echo time (TE), repetition time (TR), and case numbers, are provided in Table 1. The TMUH dataset includes a single MR source with an average resolution of 20×512×512. Data is available after registration on the NCHC website. The 2017 MICCAI WMH Segmentation Challenge dataset includes 170 sets of 3D T1-weighted and 2D multi-slice FLAIR images with WMH annotations. A total of 60 training and 110 test images were used in this challenge, with imaging data from five scanners, three vendors, and three institutes. Additional information is available at https://wmh.isi.uu.nl/ [28].

Manual annotation of the WMH

WMH were manually segmented following the STandards for ReportIng Vascular changes on nEuroimaging (STRIVE) criteria [29]. A total of 121 brain MRI series from KCGMH with WMH lesions were manually segmented using T2 FLAIR sequences by an experienced observer (C.P.C.) with in-house software from Taiwan AI Labs, previously used in other studies [30]. Manual segmentations were peer-reviewed by a second observer (Y.C.C.), who has five years of experience in clinical neuroradiology. In cases of discrepancies, the first observer corrected the segmentations in a consensus session with the second observer. The corrected segmentation by C.P.C., after peer review, serves as the reference standard. The TMUH dataset contains 505 brain MRI series with T2 FLAIR WMH segmentation labels.

AI model

We present architectural overviews of the CNN-based model in Fig. 1 and the Transformer-based model in Fig. 2. Both models use a U-shape structure, similar to the conventional U-Net [31], for medical image segmentation. The size of the input 3D MRI series is denoted as $1 \times D \times H \times W$ and the annotation of segmented masks includes the WMH foreground and background.

CNN-based model

The CNN-based model combines the conventional convolutional backbone and self-attention block. The convolutional backbone encodes spatial information into high-level features, capturing local object concepts at multiple scales. The self-attention block captures long-range contextual dependencies from global information. Consequently, this model improves upon the standard U-Net [27]. The encoder is based on the 3D ResNet-50 [26] backbone, while the decoder uses scSE blocks for self-attention and deconvolutional layers for feature upsampling. The scSE block models the interdependencies between channels [32] and recalibrates pixel-wise spatial information by projecting the features to an importance map. This enhances voxel-wise segmentation mask predictions.

Transformer-based model

The Transformer-based model is a hybrid of CNN and Transformer, built on the 3D Swin Transformer in a U-shape structure [33]. Transformers are effective at

Table 1 Scanner specifications for the KCGMH dataset

Manufacturer	Field strength	Resolutions	TR (ms)	TE (ms)	TI (ms)	Slice thickness (mm)	Flip angle	FOV	Voxel size (mm)	Number
GE Medical systems, GENESIS _ SIGNA	1.5T	256×256×20	9000	145	2200	5	90	230×230	0.82×0.82	8
GE Medical systems, SIGNA	3T	512×512×20	8000	100	2000	5	90	240×240	0.488×0.488	34
GE Medical systems, DISCOVERY MR450	1.5T	512×512×20	9000	145	2250	5	90	220×220	0.43×0.43	14
Philips Medical systems, Intera	1.5T	256×256×20	6000	110	2000	5	90	230×230	0.898×0.898	35
Siemens Medical systems, Skyra	3T	384×384×20	8000	76	2372	5	150	230×230	0.599×0.599	8
Siemens Medical systems, Skyra	3T	256×256×90	8000	90	2400	2	150	200×200	0.781×0.781	22

Abbreviations: FOV, Field-of-view; TE, Echo time; TR, Repetition time; TI, Inversion time

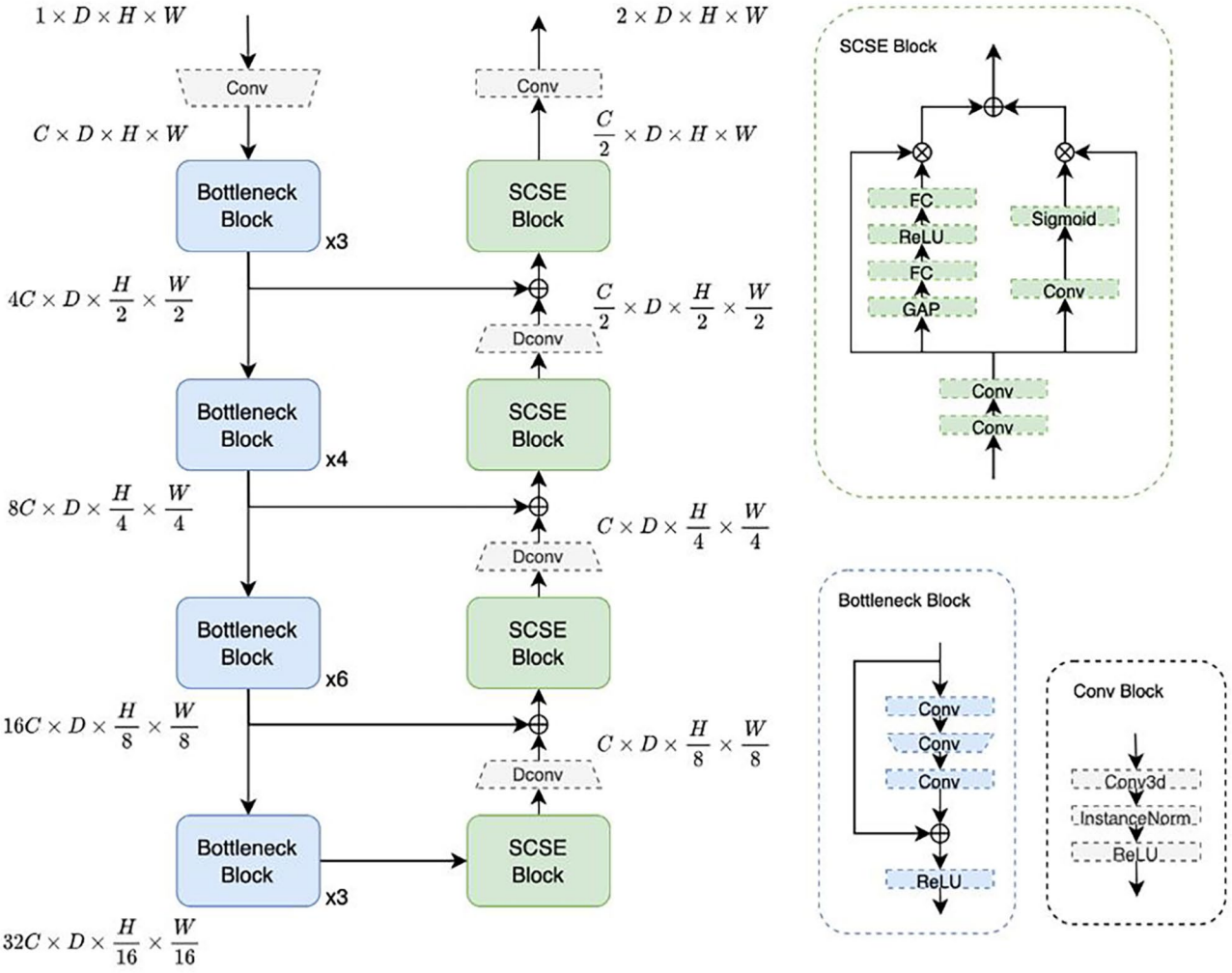

Fig. 1 Architecture of the proposed CNN-based model

aggregating global information but require large datasets for local inductive bias, which CNNs handle well. To address this, the model replaces the standard patch embedding layer with a convolutional patch embedding layer, generating overlapped patches that retain local information [14, 34–37]. In our configuration, we use a patch size of 1×4×4, followed by a linear layer to project each patch into the channel dimension C. Multiple 3D Swin Transformer blocks, with volumetric window multi-head self-attention (VW-MSA) and shifted versions (VSW-MSA), process each 3D window [27, 38]. These operations improve self-attention efficiency and control computation overhead, adapted for 3D medical images. For enabling 3D operations on medical images, we adopt a volumetric adaptation of it [38]. Specifically, at layer l, it is computed as:

$$\widehat{z}^{\,l} = VW\_MSA\,(LN(\,z^{l-1})) + z^{l-1}$$

$$z^{l} = FFN\,(LN(\,\widehat{z}^{\,l}\,)) + \widehat{z}^{\,l}$$

$$\widehat{z}^{\,l+1} = VSW\_MSA\,(LN(\,z^{l})) + z^{l}$$

$$z^{l+1} = FFN\,(LN(\,\widehat{z}^{\,l+1})) + \widehat{z}^{\,l+1}$$

Wherein LN is layer normalization, FFN is the feed-forward network, and $\widehat{z}^{\,l}$ and z^{l} denote the output features of the V(S)W-MSA module and the FFN module for block l, respectively. Patch merging layers reduce the size of embedded features in a hierarchical manner, downscaling three times with factors of (1, 2, 2), (2, 2, 2), and (2, 2, 2). Implementation details are shown in Fig. 2.

Model training

We adopt Dice loss [39] and weighted binary cross-entropy (BCE) loss, common in semantic segmentation, for training. Additionally, Dense Relative Localization (DRLoc) loss [40], originally an auxiliary self-supervised task for training Visual Transformers (VT) in computer

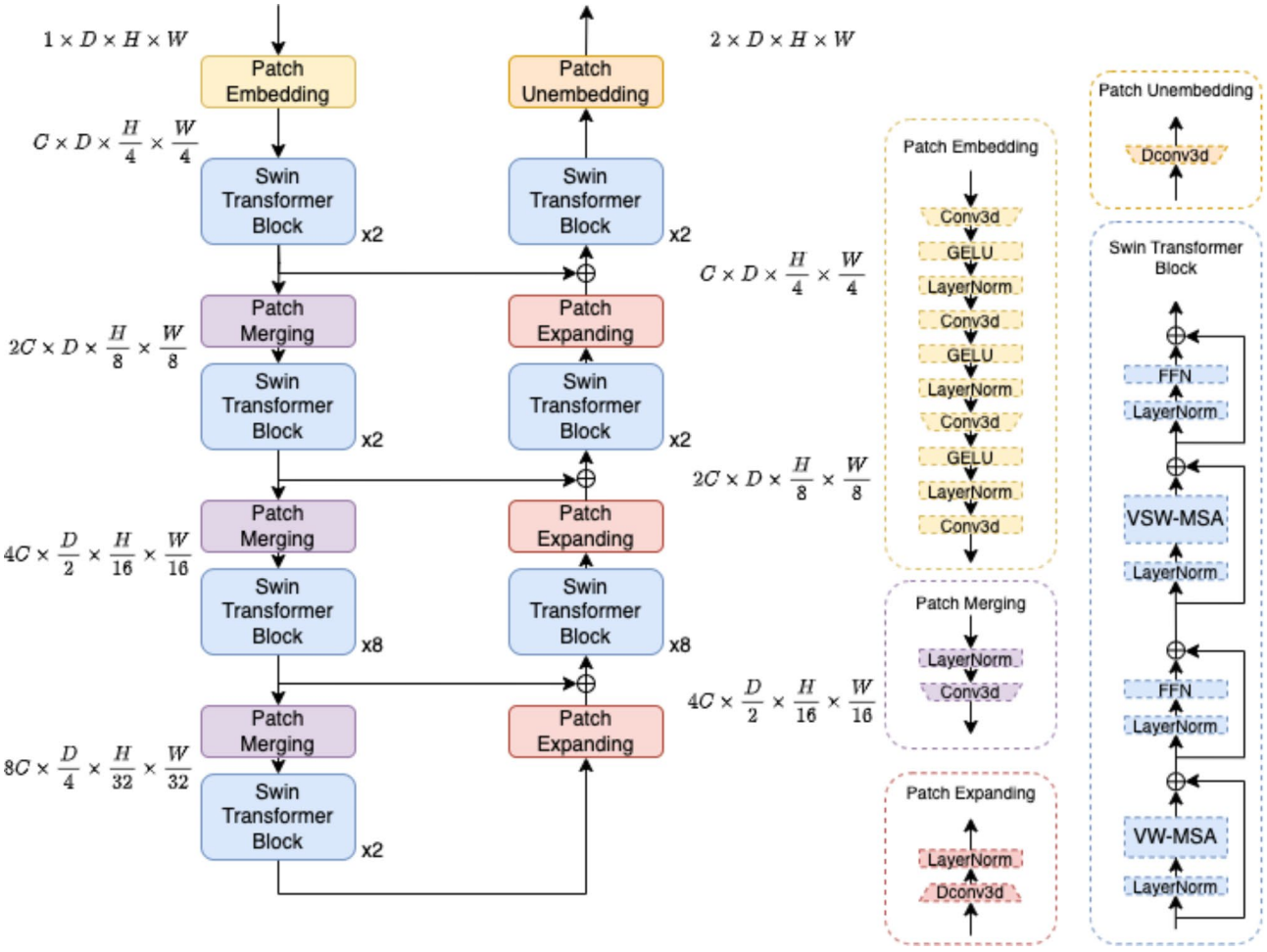

Fig. 2 Architecture of the proposed Transformer-based model

vision, is included in our Transformer-based model. DRLoc enhances the Transformer model's ability to interpret spatial relationships within an image by utilizing relative positioning data rather than relying on labeled annotations. This technique effectively supports spatial understanding in contexts where labeled data is unavailable, enabling the Transformer to capture meaningful spatial features autonomously [40]. The loss functions for the CNN-based and Transformer-based models are expressed as:

$$L_{CNN} = L_{Dice} + L_{BCE}$$

$$L_{transformer} = L_{DICE} + L_{BCE} + L_{DRLoc}$$

Pre-training with weights accelerates convergence and improves performance in high-complexity models. For the Transformer-based model, we use pre-trained 2D model weights from ImageNet [41, 42], particularly for the QKV attention and MLP layers of the Swin Transformer, as they share the same architecture. Encoder weights are reused for the decoder due to model symmetry. In contrast, the CNN-based model is randomly initialized because suitable 3D pre-trained weights are difficult to obtain.

Implementation details

The KCGMH dataset (121 series) is split into 82/20/19 (4.3:1:1) for training, validation, and testing. The TMUH dataset (505 series) is split 355/75/75 (4.7:1:1), and the mixed dataset (626 series) is split 437/95/94 (4.7:1:1). MRI inputs are normalized using min-max normalization. During training, we apply random cropping of 16×256×256 volumes and augmentations (rotation, scaling, Gaussian noise, brightness/contrast adjustment, gamma augmentation, and flipping). For validation and testing, images are resized to 256×256, and input depth is split into overlapped partitions of 16, with zero padding applied. We use an initial learning rate of 5e-5 and apply a polynomial learning rate scheduler (gamma value: 0.9). AdamW [43] is the default optimizer, with a batch size of 2 for both models. The CNN-based model

is trained for 2000 epochs, while the Transformer-based model is trained for 5000 epochs due to slower convergence. We adopt the Swin-B setting [27] with depth parameters (2, 2, 8, 2) for the Transformer-based model (Fig. 2). The Transformer model contains 86 M parameters, and the CNN model contains 87 M. All experiments were conducted on a single NVIDIA RTX 3090 GPU using Pytorch 1.8.1 [44]. We then further evaluated the models using the 2017 MICCAI WMH Segmentation Challenge dataset (60 pairs of T2-FLAIR and T1 MR images for training and 110 subjects for testing) [28].

Evaluation metrics

We evaluate performance on WMH foreground and lesion segmentation. For foreground segmentation, voxel-wise mask predictions are compared with the annotated ground truth using the Dice similarity coefficient (DSC). Lesion segmentation is evaluated by forming 3D-connected components and calculating the Intersection over Union (IoU) of predicted and ground truth lesions, with an IoU > 0.35 considered a true positive [45]. F1-Score and sensitivity (recall) are also used to assess lesion segmentation.

For the 2017 MICCAI WMH Segmentation Challenge, we evaluate using five metrics: (1) foreground DSC, (2) 95th percentile modified Hausdorff Distance (H95), (3) average volume difference (AVD), (4) lesion recall, and (5) lesion F1 score [28].

Experiments

Our experiments follow these steps:

Step (1) Train models from scratch.
Step (2) Introduce Dense Relative Localization (DRLoc).
Step (3) Utilize pre-trained model weights.
Step (4) Conduct experiments on the TMUH dataset.
Step (5) Validate models in the MICCAI challenge.

Table 2 Comparative analysis of CNN-Based and transformer-based models on the KCGMH dataset

Model	DRLoc	Pre-training	Foreground DSC	Lesion F1-Score	Lesion sensitivity
CNN-based	Without	Without	0.6128	0.4544	0.4586
	With	Without	0.6021	0.4323	0.4273
Transformer-based	Without	Without	0.6353	0.4193	0.4514
	With	Without	0.6497	0.4392	0.444
	With	With	**0.6585**	**0.5105**	**0.5363**

Training models from scratch

We began by conducting ablation studies on both CNN-based and Transformer-based models using the KCGMH dataset. Since our dataset is small compared to other large-scale image datasets, we trained the models from scratch to observe their behavior in a limited data environment.

Introduce dense relative localization (DRLoc)

To improve performance, we introduced Dense Relative Localization (DRLoc) [40]. Originally designed to address the data demands of Transformers, DRLoc is an auxiliary self-supervised task that trains alongside the primary supervised learning loss. It leverages the relative distances between embedding tokens to extract local information with minimal computational overhead, enhancing model robustness when training data is limited.

Utilizing pre-trained model weights

To further enhance the performance of the Transformer-based model, we applied 2D Swin Transformer pre-trained weights [41] from ImageNet and fine-tuned the model on the KCGMH medical image dataset. ImageNet provides a vast, labeled dataset that supports the extraction of fundamental visual features, such as edges and textures, across various imaging domains, including medical. This generalizability makes ImageNet a practical starting point for WMH segmentation in data-limited settings.

Experiments on the TMUH dataset

We extended our experiments to the TMUH dataset, which has different annotation distributions and data sources. We designed training schemes using three datasets: KCGMH, TMUH, and a combined dataset of both. Model performance was evaluated on both KCGMH and TMUH test sets.

Model validation in the MICCAI challenge

For consistency, we modified our models to support both T1 and FLAIR MRI protocols, using the training set provided in the MICCAI challenge. These adjustments maintained the same settings as our previous experiments, and we compared the performance of our models against 3D models from the challenge.

Results

Model ablation study: experiments on the KCGMH dataset

Table 2 presents results from our model ablation study, analyzed in three parts: (1) training from scratch, (2) training with DRLoc, and (3) initializing with pre-trained weights on the Transformer-based model.

Training from scratch

Without DRLoc or pre-training, the Transformer-based model achieved a foreground DSC of 0.6353, outperforming the CNN-based model's DSC of 0.6128 by nearly 2%. However, the CNN-based model outperformed in lesion segmentation by 3.5% (F1-Score: 0.4544 vs. 0.4193) (Table 2). Both models had comparable lesion sensitivity. Overall, the Transformer-based model shows potential for further improvement even with limited training data.

Training with DRLoc

Applying DRLoc [40] enhanced the Transformer model's foreground DSC by 1.5% and lesion F1-Score by 2% (Table 2), whereas it led to a decrease in the CNN-based model's performance. DRLoc therefore improves Transformer-based model performance and is used in subsequent experiments.

Pre-training on transformer-based model

Pre-training significantly improved the Transformer-based model, increasing lesion F1-Score and sensitivity by 7% and 9%, respectively, with an additional 0.9% gain in foreground DSC. Consequently, the Transformer-based model outperformed the CNN-based model across all metrics, achieving a foreground DSC of 0.6585 (CNN: 0.6128) and lesion F1-Score of 0.5105 (CNN: 0.4544) (Table 2).

Visualization

Figure 3 visualizes WMH segmentation results using the KCGMH dataset, displaying individual lesions in varied colors.

Experiments on the TMUH dataset

After configuring optimal settings on the KCGMH dataset, we trained models on three datasets: KCGMH, TMUH, and a combined dataset, then tested on both KCGMH (Table 3) and TMUH datasets (Table 4). Models trained and tested within the same dataset yielded the best performance. For instance, the Transformer-based model achieved its highest foreground DSC of 0.6585 when trained and tested on the KCGMH dataset, but performed worst (DSC: 0.308) when trained on the TMUH dataset. Mixed dataset training offered intermediate performance.

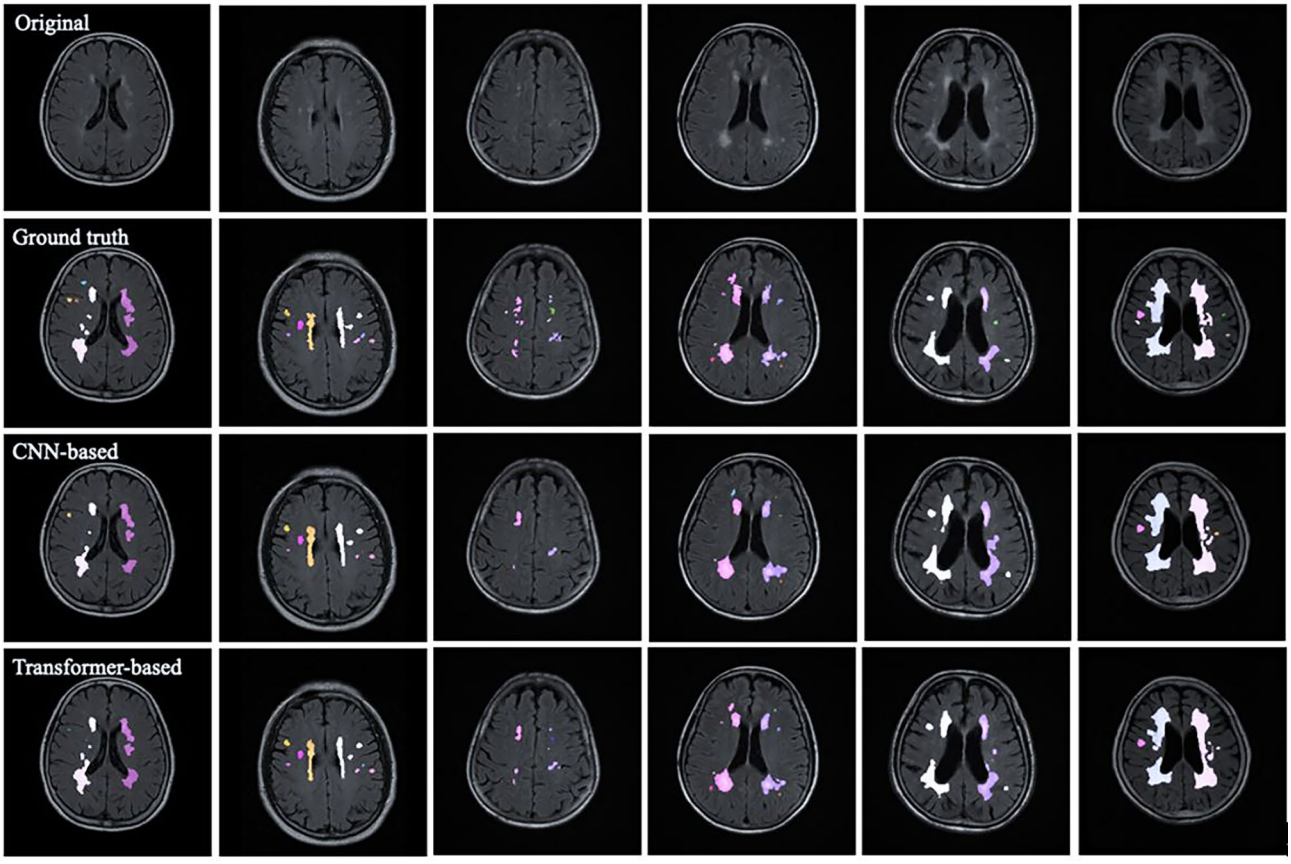

Fig. 3 WMH Segmentation Masks on FLAIR Scans from the KCGMH Dataset. (Different colors represent individual lesions. The top row shows original images, the second row shows ground truth images, the third row shows CNN-based model predictions, and the bottom row shows Transformer-based model predictions)

Table 3 Performance of the proposed models trained on varying datasets and tested on the KCGMH test set

Model	Training dataset	Foreground DSC	Lesion F1-Score	Lesion sensitivity
CNN-based	KCGMH	0.6128	0.4544	0.4586
	TMUH	0.2636	0.1835	0.152
	Mix	0.5915	0.4239	0.4232
Transformer-based	KCGMH	**0.6585**	**0.5105**	**0.5252**
	TMUH	0.308	0.2604	0.2009
	Mix	0.6415	0.47	0.4667

"Mix" denotes mixing KCGMH and TMUH training data

Table 4 Performance of the proposed models trained on varying datasets and tested on the TMUH test set

Model	Training dataset	Foreground DSC	Lesion F1-Score	Lesion sensitivity
CNN-based	KCGMH	0.3551	0.2026	0.265
	TMUH	**0.464**	0.3518	0.3491
	Mix	0.4129	0.3053	**0.3825**
Transformer-based	KCGMH	0.3686	0.2762	0.3366
	TMUH	0.4557	**0.3551**	0.3153
	Mix	0.3854	0.3016	0.3694

"Mix" denotes mixing KCGMH and TMUH training data

Table 5 Foreground segmentation statistics on the TMUH test set

Model	Training dataset	Foreground precision	Foreground recall
CNN-based	KCGMH	0.2673	0.6859
	TMUH	0.5203	0.4898
	Mix	0.3538	0.6123
Transformer-based	KCGMH	0.2896	0.6787
	TMUH	0.5304	0.4561
	Mix	0.3131	0.654

The Transformer-based model excelled across all metrics on the KCGMH test set (Table 3) but showed mixed results on the TMUH test set (Table 4). The CNN-based model had an edge in foreground segmentation on the TMUH test set, while both models were comparable in lesion segmentation (Table 4). Table 5 presents precision-recall metrics, highlighting that both models exhibited low precision and high recall when trained on KCGMH and tested on TMUH (Transformer: 0.2896/0.6787; CNN: 0.2673/0.6859). However, when both models were trained and tested on the TMUH dataset, they exhibited a different precision-recall balances (Transformer 0.5304/0.4561, CNN 0.5203/0.4898), reflecting dataset quality differences.

Performance on the MICCAI challenge

Our models were also validated on the MICCAI challenge benchmark (Table 6). Our Transformer-based model achieved top results in foreground DSC (0.79), H95 (3.71), and lesion F1-Score (0.75). Our CNN-based model attained a DSC of 0.77 and lesion F1-Score of 0.70, comparable to many benchmark models, with an H95 of 5.36 mm, outperforming all benchmark models. Both models achieved superior AVDs (CNN: 18.72, Transformer: 20.47) and lesion recall rates (CNN: 0.61, Transformer: 0.77).

Table 6 Performance evaluation of 3D models in the 2017 MICCAI WMH Segmentation Challenge

	DSC	H95(mm)	AVD(%)	Lesion Recall	Lesion F1
bigrbrain	0.77	9.46	28.04	0.78	0.71
cian	0.78	6.82	21.72	0.83	0.70
himinn	0.62	24.49	44.19	0.33	0.36
misp	0.78	11.10	19.71	0.68	0.71
neuro.ml	0.78	6.33	30.63	0.82	0.73
achilles	0.63	11.82	24.41	0.45	0.52
tignet	0.59	21.58	86.22	0.46	0.45
upc_dlmi	0.53	27.01	208.49	0.57	0.42
nic-vicorob	0.77	8.28	28.54	0.75	0.71
nus_mnndl	0.76	6.92	50.28	**0.88**	0.71
Proposed CNN-based model	0.77	5.36	**18.72**	0.61	0.7
Proposed Transformer-based model	**0.79**	**3.71**	20.47	0.77	**0.75**

Note: For each metric, the table displays the average value. Results in bold indicate the best score for each metric

Abbreviations: AVD, Average Volume Difference; DSC, Dice Similarity Coefficient; H95, 95th percentile modified Hausdorff Distance

Discussion

Experiments on the KCGMH dataset

This study explores two neural network approaches for WMH segmentation: a conventional well-established CNN-based model, widely applied in segmentation, and an emerging Transformer-based model known for its strong performance in various tasks. Through extensive experiments, we discuss practical issues related to the application of Transformers in medical imaging.

Transformer models typically require large datasets due to limited local inductive bias. DRLoc addresses this by enabling efficient Transformer training on smaller datasets, thus reducing data demands for medical imaging tasks. Pre-training further equips Transformers with essential visual pattern knowledge, decreasing training time and boosting performance. Quantitative results in Table 2 show that the Transformer-based model outperforms the CNN model across all metrics.

For qualitative performance, Fig. 3 shows WMH segmentation predictions. While both models detect larger lesions well, smaller or low-contrast lesions pose challenges. Notably, the Transformer-based model captures

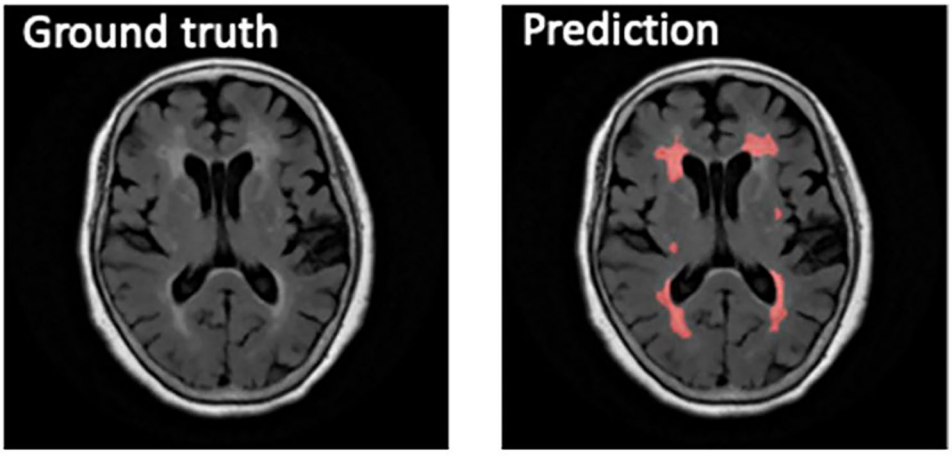

Fig. 4 Foreground segmentation predictions on the TMUH Dataset. (Left: TMUH ground truth image with incorrectly labeled WMH foregrounds as background. Right: Prediction by the CNN-based model, with the red mask indicating predicted foreground segmentation)

more small lesions, likely due to its self-attention mechanism, which captures global context.

Experiments on different datasets

We extended our analysis to three datasets: KCGMH, TMUH, and a mixed dataset. Models trained and tested on the same dataset showed optimal performance in foreground and lesion segmentation (Tables 3 and 4).

The Transformer model consistently outperformed the CNN model across all evaluation metrics on the KCGMH dataset, irrespective of whether it was trained on KCGMH, TMUH, or the mixed dataset (Table 3). However, performance varied when both models were trained and tested on the TMUH dataset. In this case, the Transformer model achieved slightly higher lesion F1-scores, whereas the CNN model performed marginally better in foreground DSC (Table 4). Notably, training on the mixed dataset produced the highest lesion sensitivity in both models when tested on TMUH, surpassing results from TMUH-only training, enhancing the model's generalizability and ability to detect diverse lesion presentations (Table 4).

Table 5 highlights an interesting finding regarding label quality differences. Models trained on the KCGMH dataset produced high recall but low precision when tested on the TMUH dataset, while models trained on TMUH achieved high precision but low recall on the same test set, indicating label quality inconsistencies. Figure 4 shows that TMUH labels often omit smaller lesions near the edges of the brain, likely impacting model predictions. Nevertheless, our models successfully segment foreground regions, even on MRI scans with coarse label quality.

Experiments on the MICCAI challenge

In the MICCAI challenge, our Transformer-based and CNN-based models achieved excellent results in foreground DSC, H95, and lesion F1-score, highlighting robust segmentation accuracy, boundary localization, and lesion detection. Our Transformer-based model led with the highest scores in foreground DSC, H95, and lesion F1-score, surpassing other competing 3D models on the benchmark. Our CNN-based model achieved the lowest AVD, affirming its strength in lesion volume estimation, while the Transformer-based model also performed well on AVD, demonstrating its high accuracy.

However, our models showed relatively lower lesion recall, which likely resulted from challenges in detecting very small or faint lesions. This limitation may stem from not applying advanced techniques aimed at improving small lesion detection, such as targeted methods, model ensemble techniques, and post-processing strategies used by models like "nic-vicorob", "nus_mnndl", "misp" and "neuroml". These methods employ multi-scale approaches, selective sampling, and data augmentation to increase sensitivity to small lesions. Adopting similar strategies in future research, such as oversampling WMH regions, incorporating multi-scale features, and refining post-processing, could further enhance recall, particularly for small lesion detection, and elevate the model's overall precision [28, 46–48]. Despite lacking specialized small lesion detection techniques, our Transformer-based model achieved excellent results, showing significant improvements across major metrics. This performance places it at the forefront of lesion segmentation, achieving the highest lesion F1 score among all models.

Limitations

While our proposed models achieve strong results in WMH segmentation, several limitations remain. First, our sample size is relatively small due to challenges in obtaining large, well-labeled medical 3D volumetric data. This limitation may affect the model's ability to learn deep, discriminative features. Expanding the dataset across multiple centers could mitigate sample size constraints and enhance generalizability.

Second, although using a single data source can streamline model performance, broader generalization requires diverse data sources. Notably, our KCGMH dataset, while limited in size, includes varied scanner types, field strengths, and resolutions, providing an initial step toward model robustness across clinical environments. Future work will incorporate additional data variability to improve applicability. Additionally, label quality in the TMUH dataset poses a challenge; higher-quality segmentations would improve the ground truth, reduce label variability, and enhance model performance. Employing consensus-based annotations and recent techniques—such as semi-automated methods like BIANCA and LST [49, 50] and advanced deep learning approaches like TrUE-Net [51]—could significantly improve label precision and model outcomes.

Third, while ImageNet pre-training offers an accessible starting point, recent studies suggest that pre-training on domain-specific medical datasets may better align model features with medical imaging needs. This trade-off highlights ImageNet's generalizability advantage while suggesting potential accuracy gains from domain-specific datasets. Limited access to large-scale, labeled 3D medical data remains a challenge. To address this, future work will explore unsupervised learning on large-scale, unlabeled 3D brain MR images to further refine our models.

Conclusion

Medical image segmentation algorithms are increasingly developed to support clinical diagnosis and treatment planning amid limited expert availability. Our study compares two modern 3D backbone networks for WMH segmentation on limited datasets. While Transformer models typically require substantial computational resources for training, their inference phase—critical for clinical application—demands significantly less. This distinction makes Transformers feasible for clinical environments with limited computational capacity, as the primary processing occurs during training. The accuracy and convergence speed of our Transformer-based model outpaced those of the CNN-based model, while both models demonstrated comparable computational demands. This study establishes a foundation for applying Transformer architectures to medical segmentation, with promising applications in resource-constrained settings where segmentation performance and efficiency are both essential.

Abbreviations

AI	Artificial intelligence
ARWMC	Age-Related White Matter Change
AVM	Arteriovenous malformation
AVF	Arteriovenous fistula
CGMH	Chang Gung Memorial Hospital
CNNs	Convolutional neural networks
DRLoc	Introduce Dense Relative Localization
DSC	Dice similarity coefficient
FLAIR	Fluid-attenuated inversion recovery
IoU	Intersection over Union
KCGMH	Kaohsiung Chang Gung Memorial Hospital
MLP	Multilayer Perceptron
MS	Multiple sclerosis
NCHC	National Center for High-Performance Computing
NLP	Natural language processing
nnFormer	Not-another transFormer
scSE	Spatial and channel squeeze & excitation
SVW-MSA	Volumetric shifted window multi-head self-attention
TAE	Transarterial embolization
TMUH	Taipei Medical University Hospital
ViT	Vision Transformer
VW-MSA	Volumetric window multi-head self-attention
WMHs	White matter hyperintensities

Acknowledgements

We thank the research participants and the research staff involved in this study. This article has not been previously published in whole or in part, in any language, except as an abstract.

Author contributions

LWC, CHL, and TYC conceived and designed the study. HYC and CYT sorted the data, prepared the figures and tables, and drafted and revised the manuscript. CHL, CPC, YCC contributed to the acquisition of the data. HYC, CYT, and LHC contributed to the analysis and interpretation of the data. HYC, LHC, TYC, and LTL provided technical support and reviewed the manuscript. LWC reviewed the manuscript and supervised the projects. All authors contributed to the article and approved the submitted version.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethical approval and consent to participate

The studies involving human participants were reviewed and approved by the Institutional Review Board of Chang Gung Memorial Hospital (IRB No. 202002026A3, passed on 19 January 2021; IRB NO. 201900483B0, passed on 10 April 2019). All patient information included in this study was anonymized and deidentified. The requirement to obtain informed consent was waived in accordance with IRB policy.

Consent for publication

Not applicable.

Competing interests

Authors TYC and LTL are executive directors of the Taiwan AI Labs. Authors HYC and LHC are employed by the Taiwan AI Labs. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 December 2023 / Accepted: 25 December 2024

Published online: 03 January 2025

References

1. Balakrishnan R, Valdés Hernández MC, Farrall AJ. Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data – a systematic review. Comput Med Imaging Graph. 2021;88:101867.
2. Li X, Zhao Y, Jiang J, Cheng J, Zhu W, Wu Z, Jing J, Zhang Z, Wen W, Sachdev PS, et al. White matter hyperintensities segmentation using an ensemble of neural networks. Hum Brain Mapp. 2022;43:929–39.
3. Iorio M, Spalletta G, Chiapponi C, Luccichenti G, Cacciari C, Orfei MD, Caltagirone C, Piras F. White matter hyperintensities segmentation: a new semi-automated method. Front Aging Neurosci. 2013;5:76.
4. Tran P, Thoprakarn U, Gourieux E, dos Santos CL, Cavedo E, Guizard N, Cotton F, Krolak-Salmon P, Delmaire C, Heidelberg D, et al. Automatic segmentation of white matter hyperintensities: validation and comparison with state-of-the-art methods on both multiple sclerosis and elderly subjects. NeuroImage: Clin. 2022;33:102940.
5. Masud M, Eldin Rashed AE, Hossain MS. Convolutional neural network-based models for diagnosis of breast cancer. Neural Comput Appl. 2022;34:11383–94.
6. Bhandari A, Koppen J, Agzarian M. Convolutional neural networks for brain tumour segmentation. Insights into Imaging. 2020;11:77.
7. Gunashekar DD, Bielak L, Hägele L, Oerther B, Benndorf M, Grosu A-L, Brox T, Zamboglou C, Bock M. Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology. Radiat Oncol. 2022;17:65.
8. Zhang C, Shu H, Yang G, Li F, Wen Y, Zhang Q, Dillenseger JL, Coatrieux JL. HIFUNet: Multi-class Segmentation of uterine regions from MR images using global Convolutional networks for HIFU surgery planning. IEEE Trans Med Imaging. 2020;39:3309–20.
9. Niyas S, Pawan S, Kumar MA, Rajan J. Medical image segmentation with 3D convolutional neural networks: a survey. Neurocomputing. 2022;493:397–413.
10. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221.
11. Lu H, Wang H, Zhang Q, Yoon SW, Won D. A 3D convolutional neural network for volumetric image semantic segmentation. Procedia Manuf. 2019;39:422–8.
12. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, Prince JL, Rueckert D, Summers RM. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE 2021;109:820–838.
13. Lei T, Wang R, Wan Y, Du X, Meng H, Nandi AK. Medical Image Segmentation Using Deep Learning: A Survey. 2020.
14. Roy AG, Navab N, Wachinger C. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In International conference on medical image computing and computer-assisted intervention. Springer; 2018:421–429.
15. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019:3146–3154.
16. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018:7132–7141.
17. Zhou H-Y, Guo J, Zhang Y, Yu L, Wang L, Yu Y. nnFormer: Interleaved Transformer for Volumetric Segmentation. 2021.
18. Hatamizadeh A, Yang D, Roth HR, Xu D. UNETR: Transformers for 3D Medical Image Segmentation. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2022:1748–1758.
19. Chang WY, Tsai MY, Lo SC. ResSaNet: A Hybrid Backbone of Residual Block and Self-Attention Module for Masked Face Recognition. In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021;2021:1468–1476.
20. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: transformers make strong encoders for medical image segmentation. arXiv Preprint arXiv: 210204306 2021.
21. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, Gao Y, Wang Q, Zhang J, Shen D. Transformers in medical image analysis: A review. arXiv 2022. arXiv preprint arXiv:220212165.
22. Yan J, Wang X, Cai J, Qin Q, Yang H, Wang Q, Cheng Y, Gan T, Jiang H, Deng J. Medical image segmentation model based on triple gate MultiLayer Perceptron. Sci Rep. 2022;12:1–14.
23. Yin XX, Sun L, Fu Y, Lu R, Zhang Y. U-Net-Based Medical Image Segmentation. J Healthc Eng 2022;2022:4189781.
24. Hassanpour N, Ghavami A. Deep learning-based Bio-medical Image Segmentation using UNet Architecture and transfer learning. arXiv Preprint arXiv:230514841 2023.
25. Azad R, Aghdam EK, Rauland A, Jia Y, Avval AH, Bozorgpour A, Karimijafarbigloo S, Cohen JP, Adeli E, Merhof D. Medical image segmentation review: the success of u-net. arXiv Preprint arXiv:221114830 2022.
26. Yu L, Yang X, Chen H, Qin J, Heng PA. Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images. In Thirty-first AAAI conference on artificial intelligence. 2017.
27. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:10012–10022.
28. Kuijf HJ, Biesbroek JM, De Bresser J, Heinen R, Andermatt S, Bento M, Berseth M, Belyaev M, Cardoso MJ, Casamitjana A. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. IEEE Trans Med Imaging. 2019;38:2556–68.
29. Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, Lindley RI, O'Brien T, Barkhof J, Benavente F. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 2013;12:822–38.
30. Huang K-C, Huang C-S, Su M-Y, Hung C-L, Ethan Tu Y-C, Lin L-C, Hwang J-J. Artificial intelligence aids cardiac image quality assessment for improving precision in strain measurements. Cardiovasc Imaging. 2021;14:335–45.
31. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention. Springer; 2015:234–241.
32. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:1492–1500.
33. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3202–3211.
34. Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:22–31.
35. Xiao T, Singh M, Mintun E, Darrell T, Dollár P, Girshick R. Early convolutions help transformers see better. Adv Neural Inf Process Syst. 2021;34:30392–400.
36. Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? Adv Neural Inf Process Syst. 2021;34:12116–28.
37. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. An image is worth 16x16 words: transformers for image recognition at scale. arXiv Preprint arXiv:201011929 2020.
38. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H. Video swin transformer. arXiv. arXiv Preprint arXiv:210613230 2021.
39. Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV). IEEE; 2016:565–571.
40. Liu Y, Sangineto E, Bi W, Sebe N, Lepri B, Nadai M. Efficient training of visual transformers with small datasets. Adv Neural Inf Process Syst. 2021;34:23818–30.
41. Xie Z, Zhang Z, Cao Y, Lin Y, Bao J, Yao Z, Dai Q, Hu H. SimMIM: A Simple Framework for Masked Image Modeling. arXiv e-prints 2021:arXiv:2111.09886.
42. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009:248–255.
43. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv Preprint arXiv:171105101 2017.
44. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in pytorch. 2017.
45. Ziabari A, Shirinifard A, Eicholtz MR, Solecki DJ, Rose DC. A two-tier convolutional neural network for combined detection and segmentation in

biological imagery. In 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE; 2019:1–5.
46. Ghafoorian M, Karssemeijer N, van Uden IW, de Leeuw FE, Heskes T, Marchiori E, Platel B. Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. Med Phys. 2016;43:6246–58.
47. Bria A, Marrocco C, Tortorella F. Addressing class imbalance in deep learning for small lesion detection on medical images. Comput Biol Med. 2020;120:103735.
48. Park G, Hong J, Duffy BA, Lee J-M, Kim H. White matter hyperintensities segmentation using the ensemble U-Net with multi-scale highlighting foregrounds. NeuroImage. 2021;237:118140.
49. Griffanti L, Zamboni G, Khan A, Li L, Bonifacio G, Sundaresan V, Jenkinson M. BIANCA (Brain Intensity AbNormality classification algorithm): a new tool for automated segmentation of white matter hyperintensities. NeuroImage. 2016;141:191–205.
50. Torres-Simon L, del Cerro-León A, Yus M, Bruña R, Gil-Martinez L, olado AM, Cuesta P. Decoding the best automated segmentation tools for vascular white matter hyperintensities in the aging brain: a clinician's guide to precision and purpose. GeroScience 2024, 1–20.
51. Sundaresan V, Zamboni G, Rothwell PM, Jenkinson M, Griffanti L. Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images. Med Image Anal. 2021;73:102184.

Publisher's note