EFFICIENT 3D TRANSFORMER WITH CLUSTER-BASED DOMAIN-ADVERSARIAL LEARNING FOR 3D MEDICAL IMAGE SEGMENTATION

Haoran Zhang[1], *Hao Chen*[2,3]

[1] Sch. of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2] Dept. of Computer Science and Engineering, The Hong Kong University of Science and Technology
[3] Dept. of Chemical and Biological Engineering, The Hong Kong University of Science and Technology

ABSTRACT

Real-world application of volumetric medical image segmentation is still challenging due to the domain shift problem and the disability to process volumetric information efficiently by existing algorithms. To address these problems, we propose a 3D Swin Transformer with a **p**yramidal **d**ownsampling strategy to process volumetric information efficiently, dubbed as PDSwin. Specifically, the improved 3D Swin Transformer includes a spatial downsampling strategy that downsamples 2D slices pyramidally according to the spatial relationship, reducing the computation complexity while providing a wider downsampled receptive field. Furthermore, we propose a cluster-based domain-adversarial learning algorithm to attenuate the domain shift problem. The algorithm generates fine-grained cluster-based domains instead of employing center-based domains, ameliorating the domain-adversarial learning performance. We evaluated our model against other competitive models on brain stroke lesion segmentation and prostate segmentation tasks. Extensive experimental results indicated that our proposed model outperforms other models, demonstrating the efficacy of our proposed method.

***Index Terms*—** Medical image segmentation, Efficient 3D Transformer, Domain-adversarial learning

1. INTRODUCTION

Automatic medical image segmentation could assist doctors in the evaluation of various diseases. However, a common challenge faced by many medical image segmentation methods is the domain shift problem, which occurs when the specific properties of medical images, such as different imaging modalities and scanners, vary between different centers. This can lead to poor segmentation performance in real-world applications. Additionally, achieving both accuracy and efficiency can be challenging in volumetric medical image segmentation tasks.

Recently, the Vision Transformer [1] has gained popularity in a variety of applications. To improve the performance, some approaches focused on reducing the complexity of global self-attention by simplifying different operations [2], while others focused on increasing the receptive field of local self-attention [3]. In the medical field, the implementation of the Transformer has made significant progress in both 2D and 3D medical image segmentation [4, 5]. However, the Transformer's high computation complexity may hinder its use in medical facilities, and it may overfit to specific domains, leading to performance degradation. To address this issue, various domain generalization (DG) methods have been proposed [6]. However, many of these DG methods utilize the available data domains coarsely, leading to a lack of fine-grained utilization of information from center-based domains.

To reduce the extra complexity of 3D Transformer and address the domain shift problem, we propose a **p**yramidally **d**ownsampled 3D Swin Transformer, dubbed as PDSwin Transformer. The PDSwin Transformer pyramidally downsamples 2D slice windows according to their spatial proximity, reducing the computation complexity and providing a wider downsampled receptive field. We also propose a cluster-based domain-adversarial learning algorithm to address the domain shift problem. Inspired by Aslani et al. [6], we construct a similar network for domain-adversarial learning and further propose a cluster-based domain generation algorithm to improve the DG ability. Extensive experiments demonstrate our proposed method's effectiveness in improving volumetric medical image segmentation.

2. METHODOLOGY

Fig. 1 illustrates the overview of our proposed method. Our method consists of two components. The first component is a U-shaped PDSwin-based network for segmentation. The second component is a domain-adversarial learning network with the cluster-based domain generation algorithm for DG. Our method aims to achieve efficient and accurate volumetric medical image segmentation on unseen domains.

2.1. Pyramidally Downsampled Transformer for Efficient Segmentation

The PDSwin Transformer is based on 3D Swin Transformer [4]. Self-attention is defined as follows:

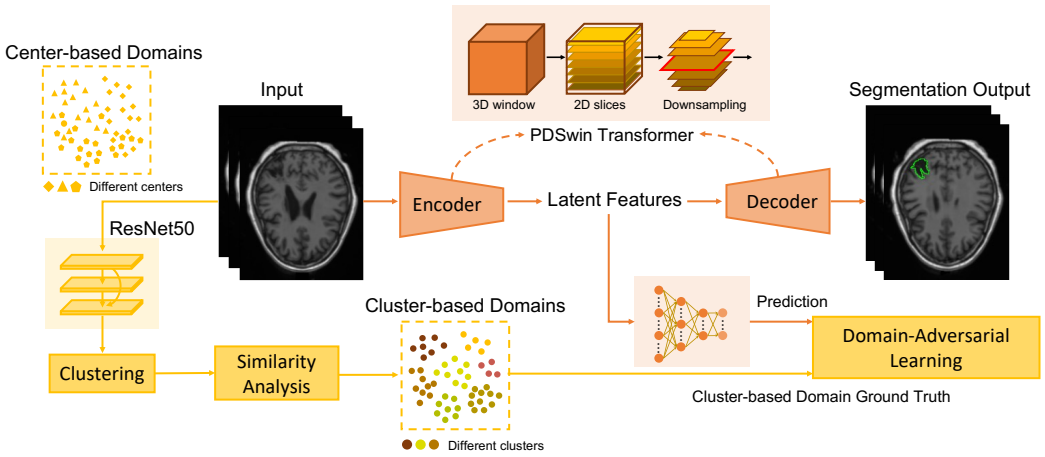


Fig. 1. An overview of our proposed method, consisting of a U-shaped PDSwin-based architecture and a cluster-based domain-adversarial learning network. The PDSwin Transformer pyramidally downsamples 2D slice windows based on their spatial proximity.

$$Attention = Softmax\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{1}$$

where $Q, K, V \in \mathbb{R}^{hwd \times C}$ denote the query, key, and value matrices, respectively. hwd and C represent the number of patches in a 3D window and the dimension of a patch, respectively. Swin Transformer [4] separates an input image into various regions, namely windows, computes local self-attention in each window, and enlarges the receptive field by shifting windows.

In the PDSwin Transformer, we assume that $Q_k \in \mathbb{R}^{hw \times C}$ denotes a query matrix of a 2D slice window with the slice index, k, such that $Q_k(i,j) = Q(i,j,k) \in \mathbb{R}^{1 \times C}$, where (i,j,k) denotes the index of a patch. $K', V' \in \mathbb{R}^{m \times C}$ denote pyramidally downsampled K and V, and m is the number of patches after downsampling. $K' = D_k(K), V' = D_k(V)$, where D_k is the pyramidal downsampling function. The self-attention is computed as follows:

$$Attention_k = Softmax\left(\frac{Q_k K'^T}{\sqrt{C}}\right)V' \tag{2}$$

where $Attention_k$ is the self-attention of slice k in a 3D window. The self-attention of the entire 3D window is defined as $Attention\,(i,j,k) = Attention_k\,(i,j)$.

In our PDSwin Transformer, we allow for 2D slice windows, p, to be involved in the computation of K' and V' for the target 2D slice window, k, if they satisfy the following requirement:

$$Distance\,(p,k) = \mid p - k \mid \leq L \tag{3}$$

where $L \in \mathbb{R}^+$ denotes a distance threshold. Theoretically, p and k should be in the same 3D window. For the convenience of programming, when the target 2D slice window k is peripheral to its 3D window, exterior 2D slice windows can also be included as long as they meet the requirement in equation (3). The downsampling level, l, of 2D slice window p is defined as $l = \frac{L - Distance(p,k)}{L}$. Given that the number of patches in 2D slice window p is hw, the number of patches in downsampled 2D slice window p is $t = \lceil hwl \rceil$. $\lceil hwl \rceil$ denotes the smallest integer greater than or equal to hwl.

We utilize average pooling for downsampling, defined as follows:

$$D_k\,(X)_{p,q} = \frac{1}{\mid \mathcal{R}_q \mid} \sum_{(i,j) \in \mathcal{R}_q} X_p(i,j) \tag{4}$$

$$D_k\,(X)_p = \left[D_k\,(X)_{p,0}\,, ..., D_k\,(X)_{p,t-1}\right]^T \tag{5}$$

where q denotes a specific patch after average pooling. $\mathcal{R}_q$ denotes the set of original patches involved in the average pooling for the patch q. $D_k\,(X)_{p,q}\,, X_p(i,j) \in \mathbb{R}^{1 \times C}$ represent the average pooled and original patch vector, respectively. $D_k(X)_p$ denotes the downsampled 2D slice window p. These processes are applied to all 2D slice windows that satisfy equation (3), resulting in the overall downsampling function $D_k\,(X)$, which is defined as:

$$D_k\,(X) = \left[D_k\,(X)_{k-L}\,, ..., D_k\,(X)_{k+L}\right]^T \tag{6}$$

By using this approach, we are able to create a pyramidally downsampled 3D window for K and V matrices, with fewer patches in 2D slice windows that are farther from the target 2D slice window k.

Theoretically, with the same computation complexity, the PDSwin Transformer supports a wider receptive field compared to the 3D Swin Transformer. The computation complexity of a 3D Swin Transformer and PDSwin Transformer is

$HWD\left(3C^2+2hwdC\right)$ and $HWD\left[C^2(2L+1)+2hwLC\right]$, respectively, where HWD denotes the number of input patches. When the computation complexity of the PDSwin Transformer is equal to that of the 3D Transformer, $\frac{2C^2+2hwC}{C^2+hwC}$ extra 2D window slices are involved in PDSwin Transformer's computation, providing an extra receptive field at a coarse-grained level.

2.2. Cluster-Based Domain-Adversarial Learning

In order to accurately utilize available domains at a fine-grained level, we propose a cluster-based domain-adversarial learning algorithm, as shown in Fig. 1 (light yellow). A ResNet50 is used to extract information from center-based domains, and K-Means clustering is applied to divide these center-based domains into cluster-based domains. We then conduct a similarity analysis among the resulting cluster-based domains, measuring the average high-dimensional vector of each cluster-based domain against the vectors of the other cluster-based domains using cosine similarity. Cluster-based domains with high similarity are merged to create clear domain demarcations.

Inspired by Aslani et al. [6], we construct a regularization network and an auxiliary loss function for DG. The regularization network consists of three perceptron layers and a softmax layer. It receives the latent features from the encoder to produce category-wise domain predictions. The auxiliary loss function is defined as follows:

$$L_{reg}(h_i, c_i) = -\sum_j h_{ij} \log c_{ij} \tag{7}$$

where c_i and h_i denote the category-wise domain prediction and one-hot encoded vector of the domain ground truth after random shuffling, respectively. c_{ij} and h_{ij} denote the components of c_i and h_i, respectively. The random shuffling confuses the encoder, impeding its ability to interpret domain information, thereby addressing the domain shift problem.

The overall loss function is defined as follows:

$$L = L_{seg} + \alpha L_{reg} \tag{8}$$

where L_{seg} denotes the soft-Dice loss function, and α is set to 0.2.

2.3. Implementation Details

All experiments were conducted using Python 3.9.12, PyTorch 1.11.0, and MONAI 0.9.1. We utilized an Nvidia GeForce RTX 3090 GPU with 24 GB of memory and a V100 GPU with 16 GB of memory during training, with a batch size of 12. Automatic mixed precision was used to accelerate training and conserve memory. Gradient accumulation was employed to maintain a consistent batch size. The initial learning rate was set to 0.0005, with exponential decay. The Adam optimizer was used for training, with a momentum of 0.9. The ResNet50 model utilized in our cluster-based domain generation algorithm was pretrained on ImageNet. Other hyperparameters were the same as those used by Hatamizadeh et al. [5]. We implemented other competitive models following their open-source codes. These models were optimized to adjust different segmentation tasks.

3. EXPERIMENTS AND RESULTS

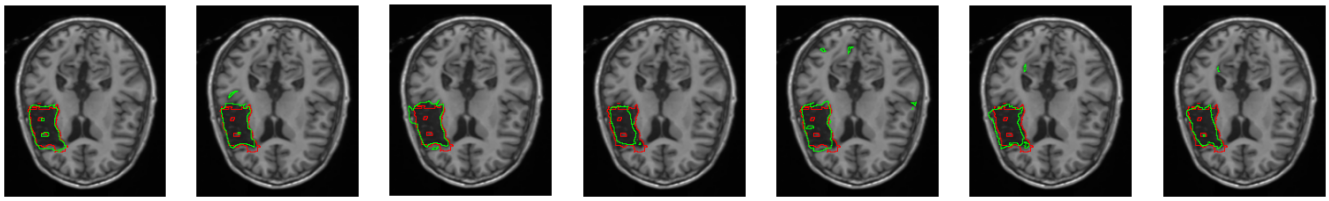
3.1. Datasets and Evaluation Metrics

Previous works [7, 8, 9] have made significant progress in brain stroke lesion and prostate segmentation. We also implemented our model in these two segmentation tasks to validate its effectiveness. For the brain stroke lesion segmentation task, the dataset contains 655 MRIs from 33 sites [10]. We divided the dataset into the training, validation, and testing sets, with 499 MRIs from 24 sites, 65 MRIs from 3 sites, and 91 MRIs from 6 sites, respectively. For the prostate segmentation task, various datasets were collected to support multiple-domains experiments, including 116 MRIs from 6 sites in total [11, 12, 13]. These datasets were divided into the training set with 79 MRIs from 3 sites, the validation set with 13 MRIs from 1 site, and the testing set with 24 MRIs from 2 sites. Cluster-based domain generation algorithm generated 84 cluster-based domains in the brain stroke lesion segmentation task and 25 cluster-based domains in the prostate segmentation task. We adopted the Dice coefficient as the evaluation metric to evaluate the model's performance.

3.2. Comparison with Other State-of-the-Art Models

We compare our method with three CNN-based and three Transformer-based methods in brain stroke lesion segmentation and prostate segmentation tasks. Center-based and cluster-based domains are utilized in training separately to evaluate the result of using different domain demarcations.

Quantitative results are reported in Table 1 and Table 2. All models with DG method improve the unseen domains segmentation accuracy, demonstrating effectiveness in applying DG method to attenuate the domain shift problem. Compared with only adopting center-based domains, the cluster-based domain generation algorithm further boosts the segmentation performance in unseen domains for all models, with an average Dice improvement of 2.61% in brain stroke lesion segmentation and 1.54% in prostate segmentation. Quantitative results also indicate that all four Transformer-based models outperform three CNN-based models, validating the capability of Transformer in volumetric segmentation. The self-attention module improves the model in abstracting long-range dependencies. Therefore, Transformer-based models are more effective in the global interpretation of image information, which is crucial in improving large-scale 3D image segmentation performance. 2D Transformer-based models, Swin Unet and Focal Unet, are worse in comparison with 3D



(1) Our method (2) UNETR[5] (3) SegResNet[14] (4) V-net[15] (5) 3D Unet[16] (6) Focal Unet[3] (7) Swin Unet[4]

Fig. 2. Qualitative comparisons of brain stroke lesion segmentation on the testing set. Red and green contours denote the ground truth and prediction, respectively.

Type	Methods	Cluster-based	Center-based	No DG
2D	Focal Unet [3]	0.5523	0.5326	0.5003
	Swin Unet [4]	0.5617	0.5340	0.4947
3D	3D Unet [16]	0.5006	0.4582	0.4153
	V-net [15]	0.5421	0.5237	0.4825
	SegResNet [14]	0.5362	0.5189	0.4746
	UNETR [5]	0.5992	0.5713	0.5394
	Our method	**0.6273**	**0.5978**	**0.5659**

Table 1. Dice scores of different models on brain stroke lesion segmentation.

Type	Methods	Cluster-based	Center-based	No DG
2D	Focal Unet [3]	0.8302	0.8138	0.7872
	Swin Unet [4]	0.8324	0.8204	0.7907
3D	3D Unet [16]	0.8157	0.8018	0.7572
	V-net [15]	0.8257	0.8085	0.7690
	SegResNet [14]	0.8232	0.8038	0.7667
	UNETR [5]	0.8420	0.8289	**0.8122**
	Our method	**0.8547**	**0.8386**	0.8119

Table 2. Dice scores of different models on prostate segmentation.

Transformer-based models, showing the importance of excavating spatial information. For two 3D Transformer-based models, Unetr and our PDSwin, the latter shows an improvement in the Dice coefficient in almost all experiments with an improvement of 1.72% in two tasks averagely.

Qualitative results of brain stroke lesion segmentation are shown in Fig. 2. The ground truth and prediction are illustrated with red and green contours, respectively. Compared with other models, our method generates more accurate prediction in the brain stroke lesion MRI segmentation task.

3.3. Evaluation on Efficiency

We evaluate the efficiency of our model with other competing models. The number of parameters and average inference time of the models in brain stroke lesion segmentation are reported in Table 3. The average inference time is measured among all volumetric cases of the testing set on a single V100 GPU mentioned in Section 2.3. For 2D Transformer-based models, Swin Unet and Focal Unet, they have a clear supremacy in efficiency compared with 3D Transformer-based models, while their segmentation accuracy is weakened due to the lack of exploiting spatial information, as mentioned in Section 3.2. For 3D models, CNN-based models are more efficient than Transformer-based models but also weakened in segmentation accuracy due to their deficiency in modeling long-range dependencies. For two 3D Transformer-based models, PDSwin-based model has an evident advance in efficiency. According to the result, our model has 81M parameters, and the average inference time is 151.1 seconds. For comparison, the pure 3D Transformer-based method, Unetr, has 92M parameters and 209.4 seconds for inference on average. With competitive segmentation accuracy illustrated in Section 3.2, our model outperforms the pure 3D Transformer-based network in efficiency and efficacy.

Type	Methods	#Params	Inference time
2D	Focal Unet [3]	59M	173.3sec
	Swin Unet [4]	57M	132.7sec
3D	3D Unet [16]	34M	188.5sec
	V-net [15]	33M	105.6sec
	SegResNet [14]	29M	108.7sec
	UNETR	92M	209.4sec
	Our method	81M	151.1sec

Table 3. Comparisons of the number of parameters and average inference time of different models on the testing set in brain stroke lesion segmentation.

4. CONCLUSION

To boost efficient and domain-generalizable medical image segmentation, a pyramidally downsampled 3D Transformer with cluster-based domain-adversarial learning is proposed in this paper. PDSwin Transformer employs a pyramidal downsampling strategy to elevate efficiency with competing segmentation accuracy. The cluster-based domain-adversarial learning algorithm increases the number of domains in training and exploits domain information at a fine-grained level. Extensive experiments on two benchmark datasets demonstrate the effectiveness of the proposed model.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki.

6. ACKNOWLEDGEMENT

This work was supported by Shenzhen Science and Technology Innovation Committee Fund (SGDX20210823103201011) and Hong Kong Innovation and Technology fund under Project ITS/028/21FP.

7. REFERENCES

[1] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[2] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6824–6835.

[3] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao, "Focal attention for long-range interactions in vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30008–30022, 2021.

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," 2021.

[5] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[6] Shahab Aslani, Vittorio Murino, Michael Dayan, Roger Tam, Diego Sona, and Ghassan Hamarneh, "Scanner invariant multiple sclerosis lesion segmentation from mri," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 781–785.

[7] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2018.

[8] Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng, "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 447–456.

[9] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng Ann Heng, "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[10] Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al., "A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms," *Scientific data*, vol. 9, no. 1, pp. 1–12, 2022.

[11] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Philips S, Maffitt D, Pringle M, Tarbox L, and Prior F, "NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures," Apr. 2020.

[12] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau, "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review," *Computers in biology and medicine*, vol. 60, pp. 8–31, 2015.

[13] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al., "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[14] Andriy Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

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

[16] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al., "U-net: deep learning for cell counting, detection, and morphometry," *Nature methods*, vol. 16, no. 1, pp. 67–70, 2019.