CATS v2: Hybrid encoders for robust medical segmentation

Hao Li[a], Han Liu[b], Dewei Hu[a], Xing Yao[b], Jiacheng Wang[b], and Ipek Oguz[a,b]

[a]Department of Electrical and Computer Engineering, Vanderbilt University
[b]Department of Computer Science, Vanderbilt University

ABSTRACT

Convolutional Neural Networks (CNNs) exhibit strong performance in medical image segmentation tasks by capturing high-level (local) information, such as edges and textures. However, due to the limited field of view of convolution kernels, it is hard for CNNs to fully represent global information. Recently, transformers have shown good performance for medical image segmentation due to their ability to better model long-range dependencies. Nevertheless, transformers struggle to capture high-level spatial features as effectively as CNNs. A good segmentation model should learn a better representation from local and global features to be both precise and semantically accurate. In our previous work, we proposed CATS, which is a U-shaped segmentation network augmented with transformer encoder. In this work, we further extend this model and propose CATS *v2* with hybrid encoders. Specifically, hybrid encoders consist of a CNN-based encoder path paralleled to a transformer path with a shifted window, which better leverage both local and global information to produce robust 3D medical image segmentation. We fuse the information from the convolutional encoder and the transformer at the skip connections of different resolutions to form the final segmentation. The proposed method is evaluated on three public challenge datasets: Beyond the Cranial Vault (BTCV), Cross-Modality Domain Adaptation (CrossMoDA) and task 5 of Medical Segmentation Decathlon (MSD-5), to segment abdominal organs, vestibular schwannoma (VS) and prostate, respectively. Compared with the state-of-the-art methods, our approach demonstrates superior performance in terms of higher Dice scores. Our code is publicly available at https://github.com/MedICL-VU/CATS.

Keywords: Convolutional neural network, Transformer, Hybrid encoder, Medical image segmentation

1. INTRODUCTION

In recent years, deep learning (DL) has shown excellent performance in many medical image segmentation tasks.[1] As a fundamental unit of DL, convolutional neural networks (CNNs) are widely used for segmentation due to their ability to learn complex patterns and structures from medical datasets. By hierarchically learning parameters using both linear and non-linear layers, CNNs leverage both local and global information from images to predict segmentations. For instance, U-Net[2] is a popular architecture specifically designed for biomedical image segmentation. This U-shaped network consists of an encoder and a decoder, interconnected by skip connections. These connections ensure that high-resolution features are combined with upsampled low-resolution features to facilitate precise segmentation. Furthermore, variants of U-Net have demonstrated state-of-the-art performance across various medical image segmentation tasks and different imaging modalities.[3–9] However, due to the local receptive field of convolution kernels, convolutional encoders have limitations in modeling long-range dependencies and potentially missing out on global context in medical images.

Inspired by the success of the Vision Transformer (ViT),[10] transformers have recently been adapted to the medical imaging field to produce high-quality segmentation.[11–13] These transformer-based methods process an input image/patch as a sequence of subpatches, rather than analyzing the entire input at once. With this property, the primary advantage of transformers is their ability to model long-range dependencies using the self-attention mechanism and to interact with all pixels in the image, in contrast to CNNs which possess a localized field of view. This global perspective is especially valuable in medical image segmentation, where contextual information from distant parts of the image can be important. However, ViT is computationally

Further author information: (Send correspondence to Hao Li)
Hao Li: E-mail: hao.li.1@vanderbilt.edu

intensive and struggles to capture local information, especially for high-resolution medical data. As a variant of the ViT, the Swin Transformer[14] has shown good performance by computing representations hierarchically within shifted windows instead of applying self-attention to the entire image. Compared to ViT, the Swin Transformer reduces computational redundancies using the shifted window scheme, and it has been utilized in medical applications to produce robust segmentations from high-resolution medical data.[15–18] In addition to preserving global information, the shifted window approach also enhances the capture of local details. Given the importance of precise segmentation of anatomical structures and pathological regions in medical imaging, the ability to focus on fine-grained details is particularly advantageous for tasks like tumor and multi-organ segmentation.[15–18]

Although the shifted window approach is effective, it may still not match the local specificity of a carefully designed CNN for certain medical image segmentation tasks, because fine-level details can be of paramount importance in medical imaging. Thus, a hybrid approach that combines the strengths of both CNN and transformer might provide an optimal solution.[19–21] This raises the question: could hybrid encoders incorporating Swin Transformers enhance the current segmentation networks used for 3D medical image segmentation?

In this work, we introduce a 3D segmentation network with hybrid encoders named CATS *v2*. This is an improved version of our previous work, CATS (complementary CNN and transformer encoders for segmentation),[20] and offers better performance. In particular, we replace the ViT with the Swin Transformer, which is used as an additional independent encoder in a U-shaped CNN. The multi-scale features extracted from the Swin Transformer are fused with the features from the CNN and then delivered to the CNN-based decoder for segmentation. We evaluate the proposed methods on three different segmentation tasks, including abdominal organs, vestibular schwannoma (VS), and prostate, where large inter-subject variations are present. We compare our model to state-of-the-art models on three public datasets. The better performance of the proposed method in terms of Dice scores indicates that Swin Transformer improves the segmentation ability of existing segmentation networks with hybrid encoders. Moreover, our method has the potential to serve as a backbone for recent methods[21–25] based on the Segment Anything Model (SAM[26]) in the field of medical image segmentation.

2. METHODS

2.1 Framework overview

Fig. 1 (a) shows the proposed segmentation network with hybrid encoders. Our model consists of two encoder paths: a CNN path and a transformer path with shifted window. The CNN-based encoder progressively encodes information using convolution and downsampling operations. On the Transformer path, the input images pass through the patch partition layer to reduce the dimension and visualize high-level features by a convolution operation and are then fed into the transformer blocks. The information from both paths is fused at each level using addition operations, and this combined information is delivered to the CNN-based decoder to predict the final segmentation.

2.2 Swin Transformer encoder

The proposed Swin Transformer encoder is adopted from.[14,15] Specifically, the input of the Swin Transformer encoder is a 3D image, and a patch partition layer is applied to create a sequence of 3D patches/tokens with a given patch size. However, unlike ViT that flattens these patches and feeds them directly into the Transformer, non-overlapping local windows are created for efficient patch interaction modeling. Each local window goes through a linear projection layer to transform it into a sequence of token vectors. The transformed vectors are then processed by the self-attention mechanism of the transformer. Our encoder has four Swin blocks and each contains two successive transformer layers, i.e., regular window multi-head self-attention (W-MSA) and shifted window MSA (SW-MSA), which are shown in Fig. 1 (a).

Fig. 1 (b) demonstrates the shifted window scheme for subsequent transformer layers. In the layer l (W-MSA), we evenly partition the patch into subregions with same window size at each dimension. In the subsequent layer, $l+1$, the partitioned window regions are shifted by half of window size. The position of the windows is shifted to allow the model to gradually increase its receptive field and incorporate a more global context into its representations. To preserve the hierarchical structure of the encoder, a patch merging layer is employed at

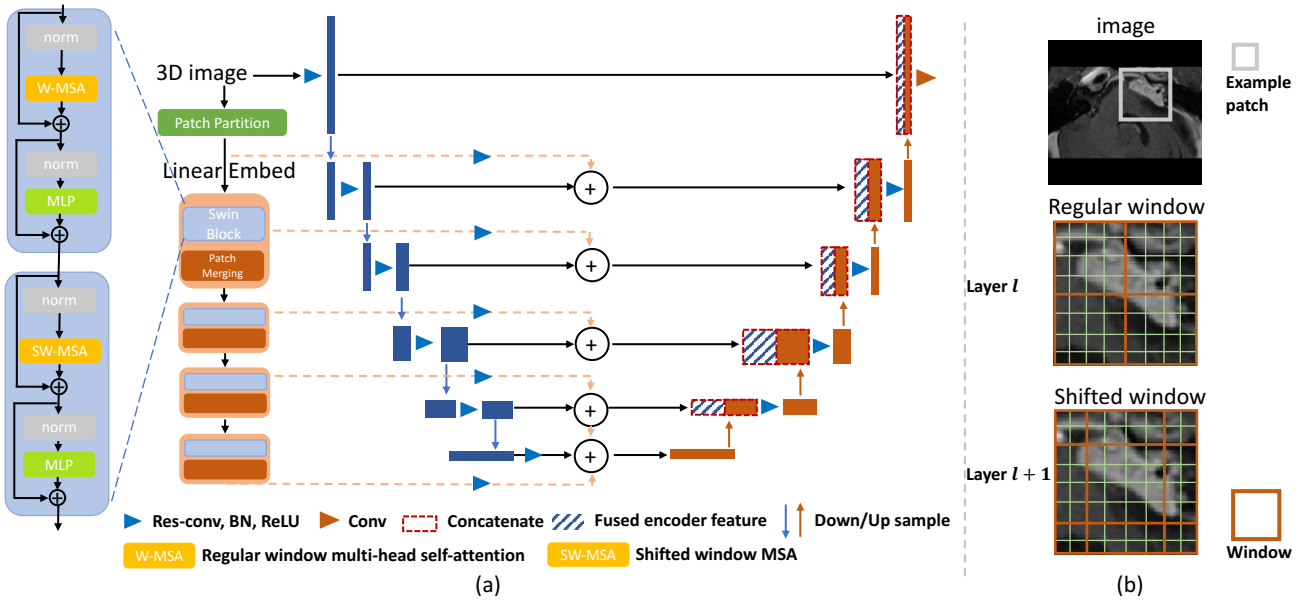

Figure 1. **(a)** Proposed network architecture. **(b)** 2D illustrations of shifted window where self-attention is only computed within each non-overlapping local window. Note that the patch sizes vary.

the end of each stage. This reduces the resolution of feature representations by a factor of 2, thereby decreasing the complexity and increasing the efficiency of the model. Following Hatamizadeh et al.,[15] the embedding layer reduces the dimension of its input by half. Note that the linear projection layer enables the model to efficiently handle high-resolution inputs by reducing the dimensionality.

2.3 Convolutional neural network architecture

Fig. 1 (a) also shows the proposed CNN, which is adapted from the 3D U-Net and its variants.[27,28] Max-pooling and deconvolution operations are employed for down-sample and up-sample, respectively. The feature maps from the highest level are sent directly to the decoder, while feature maps from the lower levels are combined with encoded information from the Swin Transformer encoder path via addition. This fused information is then delivered to the decoder using skip connections, following the pattern of the 3D U-Net[27] to produce the final segmentation.

2.4 Datasets

We use three publicly available datasets in our experiments.

- The **BTCV**[29] dataset contains 30/20 subjects with abdominal CT images for training/testing, with 13 different organs labeled by experts. The results are obtained from the official leaderboard.
- **CrossModa**[30] has 105 contrast-enhanced T1-weighted MRIs with manual labels for vestibular schwannomas (VS). We split the dataset into 55/20/30 for training/validation/testing.
- **MSD-5**[31] consists of 32 MRIs with manual prostate labels. 2 MRIs in validation were excluded due to the wrong labels being provided in the public dataset. We use this dataset in a 5-fold cross-validation framework, and follow the setting in nnUnet.[32]

Dice score, average surface distance (ASD) and 95-percent Hausdorff distance (HD95) are used as evaluation metrics. The details of preprocessing steps for all datasets can be found in the original CATS paper.[20]

Table 1. Mean Dice scores in BTCV dataset. Bold numbers denote the highest Dice scores. The results of TransUNet are directly copied from.[12] The experiments follow the public pipeline of Swin UNETR.[15] The organs from left to right are: spleen, right and left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right and left adrenal gland, and overall average. Bold numbers indicate the best performance. The results can be found on the official leaderboard.

Method	Spl	RKid	LKid	Gall	Eso	Liv	Sto
TransUNet[12]	85.1	77.0	81.9	63.1	-	94.1	75.6
UNETR[13]	93.4	85.5	87.6	61.9	74.7	95.7	76.8
Swin UNETR[15]	**95.9**	87.8	92.9	65.7	77.2	96.5	83.3
CATS	95.8	**90.2**	**93.4**	65.9	77.1	**96.8**	83.0
CATS *v2*	94.8	87.1	93.2	**70.7**	**78.1**	96.7	**85.8**
	Aor	IVC	Veins	Pan	RAG	LAG	Avg.
TransUNet[12]	87.2	-	-	55.9	-	-	77.5
UNETR[13]	85.2	77.2	69.8	61.5	64.4	59.4	76.9
Swin UNETR[15]	85.5	82.8	75.1	72.5	**74.0**	**72.0**	81.6
CATS	**88.6**	**83.1**	76.9	73.8	70.2	62.6	81.4
CATS *v2*	88.0	82.5	**77.0**	**76.1**	72.2	66.3	**82.2**

2.5 Implementation details

We followed the implementation settings in CATS[20] for our experiments for a fair comparison. Briefly, we normalized the image intensity to range [0, 1]. The constant learning rate was set to 0.0001. Training batch size was 2 for all experiments which are conducted on Pytorch, MONAI and an NVidia Titan RTX GPU.

3. RESULTS

3.1 BTCV results

The quantitative and qualitative results of BTCV dataset are shown in Tab. 1 and Fig. 2, respectively. The compared methods include TransUNet,[12] UNETR,[13] Swin UNETR,[15] CATS,[20] and the proposed CATS *v2*. Briefly, UNETR[13] is composed of a ViT encoder and a CNN decoder, while Swin UNETR replaces the ViT with a Swin encoder. Similarly, CATS[20] is built upon the 3D U-Net[27] and integrates a ViT encoder. The proposed CATS *v2* employs a Swin encoder as the upgrade.

From Tab. 1, the proposed CATS *v2* achieves the best overall performance among the state-of-the-art compared methods (the 'Avg.' column). In the comparison between Swin UNETR and proposed CATS *v2*, we observe the improvements in 8 out of 13 organs when a CNN encoder is integrated. Furthermore, the proposed CATS *v2* outperforms original CATS in 7 out of 13 organs, with larger improvements observed in organs of smaller volume, such as the gallbladder, and the right and left adrenal glands. These improvements suggest that the Swin encoder could further refine the local details. Fig. 2 shows qualitative results, with major differences highlighted by orange arrows. Compared to the Swin UNETR and the original CATS, our proposed model produces smoother results.

Table 2. Quantitative results in CrossMoDA dataset, presented as *mean*(*std.dev.*). Bold numbers indicate the best performance.

Method	Dice	ASD	HD95
2.5D CNN[33]	0.856 (1.000)	0.69 (1.20)	3.5 (5.2)
TransUNet[12]	0.792 (0.234)	7.86 (27.6)	12 (31)
UNETR[13]	0.772 (0.139)	7.95 (14.2)	26 (43)
CATS[20]	0.873 (0.088)	**0.48 (0.63)**	2.6 **(3.6)**
CATS *v2*	**0.886 (0.076)**	**0.48** (0.79)	**2.4** (4.0)

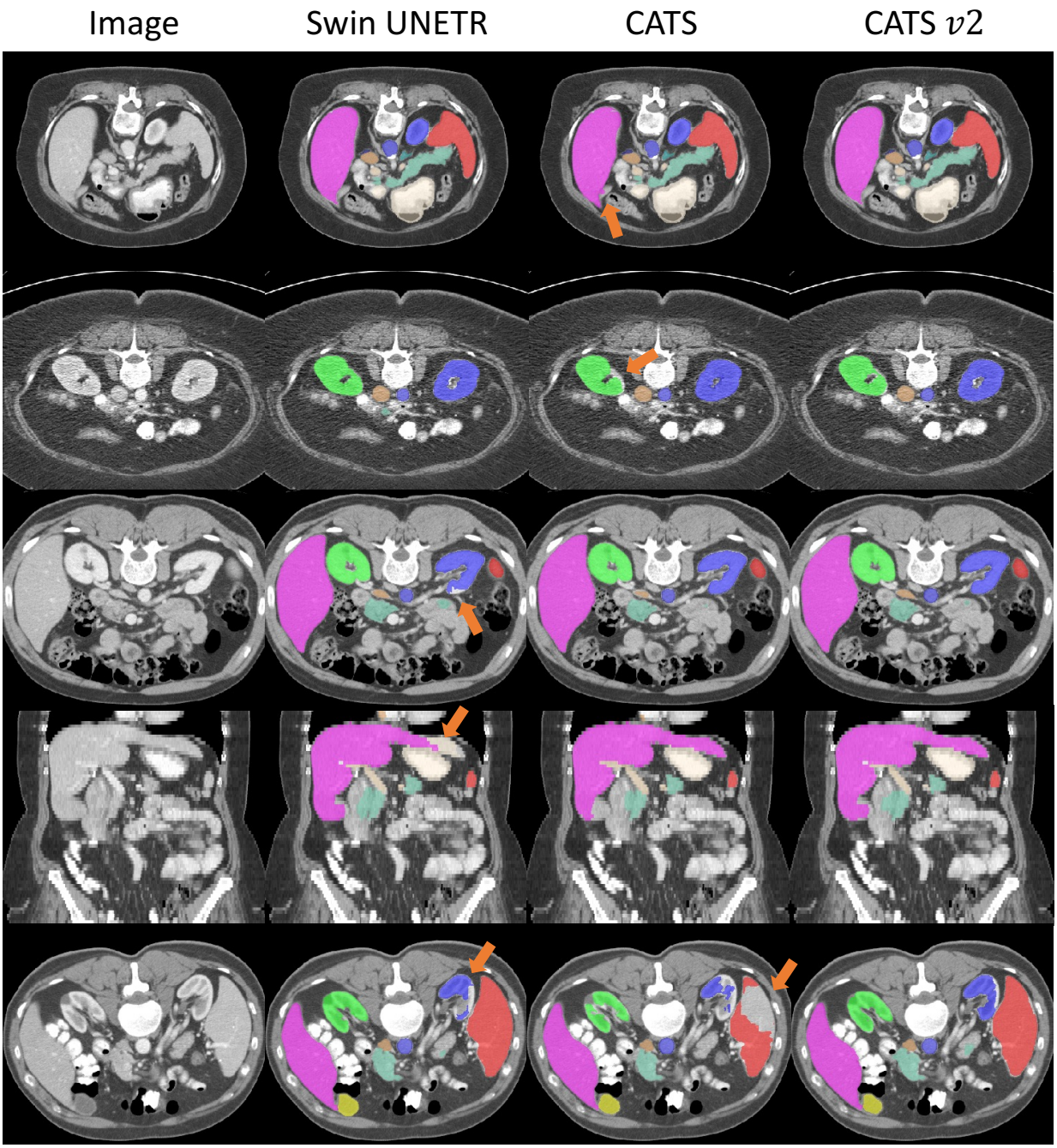

Figure 2. Qualitative results in BTCV. Some major differences are highlighted by orange arrows.

3.2 CrossMoDA results

The quantitative results for the CrossMoDA dataset are presented in Tab. 2. We compare the models against a 2.5D CNN model,[33] which was specifically designed to segment VS from MRIs characterized by substantial discrepancies between in-plane resolution and slice thickness, which is a common feature of this dataset. We observe that this CNN-only network performs better than the transformer-based encoders[12,13] for this task. The original CATS[20] model outperformed the 2.5D CNN. With subsequent enhancements, our updated CATS *v2* model further refined the quality of segmentation, delivering the highest performance in terms of Dice score. Fig. 3 shows the qualitative results of VS segmentation. While the original CATS model undersegments the

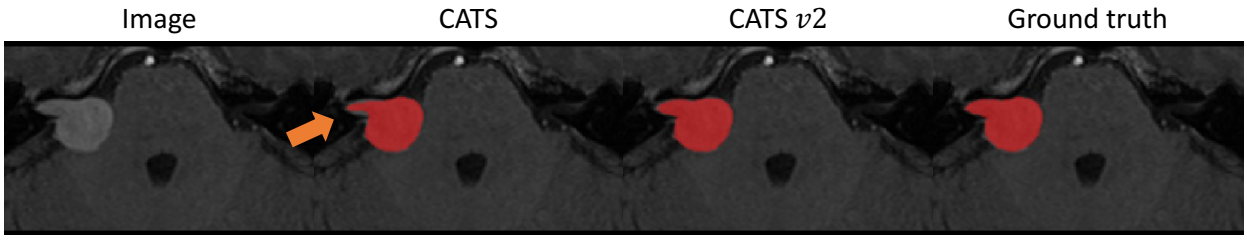

Figure 3. Qualitative results in CrossMoDA. Local segmentation errors are highlighted with arrows.

VS (marked by arrow), the proposed CATS *v2* effectively compensates for this limitation and produces robust results that align more closely with the ground truth segmentations.

3.3 MSD-5 results

We compared the nnUnet,[32] TransFuse[19] and CATS[20] to our proposed method for the prostate segmentation task in Tab. 3. CATS *v2* has the highest Dice scores on all labels, i.e., both the peripheral zone (PZ) and the transition zone (TZ). This dataset was chosen because of the inherent challenge in segmenting two closely adjoined regions that exhibit considerable inter-subject variability. The qualitative improvements between original CATS and CATS *v2* are shown in Fig. 4. A more robust segmentation is produced by the proposed method by correcting the false positives.

4. DISCUSSION AND CONCLUSION

In this work, we introduce CATS *v2*, which is a segmentation network with hybrid encoders, specifically, a U-shaped CNN complemented with a Swin Transformer. We evaluated our proposed methods on three public datasets that present large inter-subject variations. Our proposed model outperforms state-of-the-art models on each task. Relative to the original CATS, the Swin Transformer is able to further enhance the segmentation ability of the encoder. However, we observe inconsistent improvements in the BTCV dataset, indicating that one encoder may dominate the results. Exploration of other fusion strategies to overcome this issue remains as future work. In addition, due to the use of hybrid encoders as well as deeper architecture design, our proposed network might require slightly more computational resources than the original CATS. In the future work, we aim to design a light-weight model for 3D medical image segmentation.

Table 3. Mean Dice scores in MSD-5 dataset. PZ and TZ denote the peripheral zone and the transition zone, respectively. Bold numbers indicate the best performance.

Method	PZ	TZ	Avg.
2D nnUnet[32]	0.6285	0.8380	0.7333
3D nnUnet[32]	0.6663	0.8410	0.7537
TransFuse[19]	0.6738	0.8539	0.7639
CATS[20]	0.7136	0.8618	0.7877
CATS *v2*	**0.7356**	**0.8713**	**0.8034**

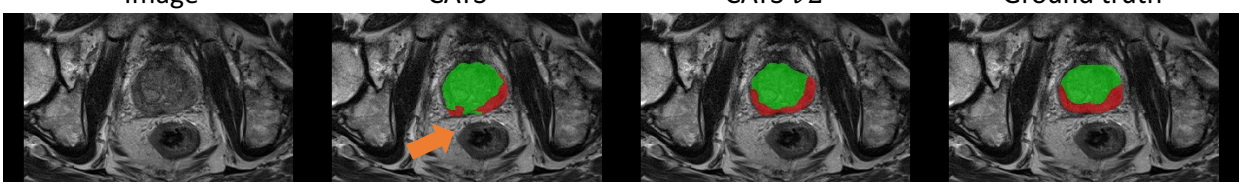

Figure 4. Qualitative results in MSD-5. Local segmentation errors are highlighted with arrows. Red and green labels denote the peripheral zone (PZ) and the transition zone (TZ), respectively.

Acknowledgements. This work was supported, in part, by NIH grant U01-NS106845, NIH grant R01-NS094456, and NSF grant 2220401.

REFERENCES

[1] Liu, H., Hu, D., Li, H., and Oguz, I., "Medical image segmentation using deep learning," in [*Machine Learning for Brain Disorders*], 391–434, Springer (2023).

[2] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*], 234–241, Springer (2015).

[3] Zhang, H., Li, H., and Oguz, I., "Segmentation of new ms lesions with tiramisu and 2.5 d stacked slices," *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure* **61** (2021).

[4] Li, H., Zhang, H., Johnson, H., Long, J. D., Paulsen, J. S., and Oguz, I., "Mri subcortical segmentation in neurodegeneration with cascaded 3d cnns," in [*Medical Imaging 2021: Image Processing*], **11596**, 236–243, SPIE (2021).

[5] Li, H., Zhang, H., Johnson, H., Long, J. D., Paulsen, J. S., and Oguz, I., "Longitudinal subcortical segmentation with deep learning," in [*Medical Imaging 2021: Image Processing*], **11596**, 73–81, SPIE (2021).

[6] Hu, D., Cui, C., Li, H., Larson, K. E., Tao, Y. K., and Oguz, I., "Life: a generalizable autodidactic pipeline for 3d oct-a vessel segmentation," in [*Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*], 514–524, Springer (2021).

[7] Li, H., Zhu, Q., Hu, D., Gunnala, M. R., Johnson, H., Sherbini, O., Gavazzi, F., D'Aiello, R., Vanderver, A., Long, J. D., et al., "Human brain extraction with deep learning," in [*Medical Imaging 2022: Image Processing*], **12032**, 369–375, SPIE (2022).

[8] Li, H., Liu, H., Hu, D., Wang, J., Johnson, H., Sherbini, O., Gavazzi, F., D'Aiello, R., Vanderver, A., Long, J., et al., "Self-supervised test-time adaptation for medical image segmentation," in [*International Workshop on Machine Learning in Clinical Neuroimaging*], 32–41, Springer (2022).

[9] Liu, H., Fan, Y., Li, H., Wang, J., Hu, D., Cui, C., Lee, H. H., Zhang, H., and Oguz, I., "Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 444–453, Springer (2022).

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929* (2020).

[11] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., and Fu, H., "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873* (2022).

[12] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306* (2021).

[13] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D., "Unetr: Transformers for 3d medical image segmentation," in [*Proceedings of the IEEE/CVF winter conference on applications of computer vision*], 574–584 (2022).

[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 10012–10022 (2021).

[15] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., and Xu, D., "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," *arXiv preprint arXiv:2201.01266* (2022).

[16] Peiris, H., Hayat, M., Chen, Z., Egan, G., and Harandi, M., "A robust volumetric transformer for accurate 3d tumor segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 162–172, Springer (2022).

[17] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M., "Swin-unet: Unet-like pure transformer for medical image segmentation," in [*European conference on computer vision*], 205–218, Springer (2022).

[18] Zhou, H.-Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., and Yu, Y., "nnformer: Volumetric medical image segmentation via a 3d transformer," *IEEE Transactions on Image Processing* (2023).

[19] Zhang, Y., Liu, H., and Hu, Q., "Transfuse: Fusing transformers and cnns for medical image segmentation," in [*Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*], 14–24, Springer (2021).

[20] Li, H., Hu, D., Liu, H., Wang, J., and Oguz, I., "Cats: Complementary cnn and transformer encoders for segmentation," in [*2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*], 1–5, IEEE (2022).

[21] Li, H., Liu, H., Hu, D., Wang, J., and Oguz, I., "Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models," *arXiv preprint arXiv:2310.19721* (2023).

[22] Li, H., Liu, H., Hu, D., Wang, J., and Oguz, I., "Assessing test-time variability for interactive 3d medical image segmentation with diverse point prompts," *arXiv preprint arXiv:2311.07806* (2023).

[23] Yao, X., Liu, H., Hu, D., Lu, D., Lou, A., Li, H., Deng, R., Arenas, G., Oguz, B., Schwartz, N., et al., "False negative/positive control for sam on noisy medical images," *arXiv preprint arXiv:2308.10382* (2023).

[24] Wang, J., Li, H., Hu, D., Tao, Y. K., and Oguz, I., "Novel oct mosaicking pipeline with feature-and pixel-based registration," *arXiv preprint arXiv:2311.13052* (2023).

[25] Zhang, Y., Shen, Z., and Jiao, R., "Segment anything model for medical image segmentation: Current applications and future directions," *arXiv preprint arXiv:2401.03495* (2024).

[26] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al., "Segment anything," *arXiv preprint arXiv:2304.02643* (2023).

[27] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O., "3d u-net: learning dense volumetric segmentation from sparse annotation," in [*International conference on medical image computing and computer-assisted intervention*], 424–432, Springer (2016).

[28] Li, H., Hu, D., Zhu, Q., Larson, K. E., Zhang, H., and Oguz, I., "Unsupervised cross-modality domain adaptation for segmenting vestibular schwannoma and cochlea with data augmentation and model ensemble," in [*International MICCAI Brainlesion Workshop*], 518–528, Springer (2021).

[29] Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., and Klein, A., "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in [*Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*], **5**, 12 (2015).

[30] Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al., "Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation," *Medical Image Analysis* **83**, 102628 (2023).

[31] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al., "The medical segmentation decathlon," *Nature communications* **13**(1), 4128 (2022).

[32] Isensee, F., Jäger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H., "Automated design of deep learning methods for biomedical image segmentation," *arXiv preprint arXiv:1904.08128* (2019).

[33] Shapey, J., Wang, G., Dorent, R., Dimitriadis, A., Li, W., Paddick, I., Kitchen, N., Bisdas, S., Saeed, S. R., Ourselin, S., et al., "An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced t1-weighted and high-resolution t2-weighted mri," *Journal of neurosurgery* **134**(1), 171–179 (2019).