SwinMM: Masked Multi-view with Swin Transformers for 3D Medical Image Segmentation

Yiqing Wang[1], Zihan Li[2], Jieru Mei[3], Zihao Wei[1,7], Li Liu[4], Chen Wang[5], Shengtian Sang[6], Alan L. Yuille[3], Cihang Xie[4], and Yuyin Zhou[4(✉)]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] University of Washington, Seattle, USA
[3] The Johns Hopkins University, Baltimore, USA
[4] University of California, Santa Cruz, Santa Cruz, USA
yzhou284@ucsc.edu
[5] Tsinghua University, Beijing, China
[6] Stanford University, Stanford, USA
[7] University of Michigan, Ann Arbor, Ann Arbor, USA

Abstract. Recent advancements in large-scale Vision Transformers have made significant strides in improving pre-trained models for medical image segmentation. However, these methods face a notable challenge in acquiring a substantial amount of pre-training data, particularly within the medical field. To address this limitation, we present **M**asked **M**ulti-view with **Swin** Transformers (**SwinMM**), a novel multi-view pipeline for enabling accurate and data-efficient self-supervised medical image analysis. Our strategy harnesses the potential of multi-view information by incorporating two principal components. In the pre-training phase, we deploy a masked multi-view encoder devised to concurrently train masked multi-view observations through a range of diverse proxy tasks. These tasks span image reconstruction, rotation, contrastive learning, and a novel task that employs a mutual learning paradigm. This new task capitalizes on the consistency between predictions from various perspectives, enabling the extraction of hidden multi-view information from 3D medical data. In the fine-tuning stage, a cross-view decoder is developed to aggregate the multi-view information through a cross-attention block. Compared with the previous state-of-the-art self-supervised learning method Swin UNETR, SwinMM demonstrates a notable advantage on several medical image segmentation tasks. It allows for a smooth integration of multi-view information, significantly boosting both the accuracy and data-efficiency of the model. Code and models are available at https://github.com/UCSC-VLAA/SwinMM/.

Y. Wang, Z. Li, J. Mei and Z. Wei—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_47.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 486–496, 2023.
https://doi.org/10.1007/978-3-031-43898-1_47

1 Introduction

Medical image segmentation is a critical task in computer-assisted diagnosis, treatment planning, and intervention. While large-scale transformers have demonstrated impressive performance in various computer vision tasks [7,10,15], such as natural image recognition, detection, and segmentation [5,16], they face significant challenges when applied to medical image analysis. The primary challenge is the scarcity of labeled medical images due to the difficulty in collecting and labeling them, which requires specialized medical knowledge and is time-consuming [12,23,25]. The second challenge is the ability to identify sparse and obscure patterns in medical images, including blurred and dim images with small segmentation targets. Hence, it is imperative to develop a precise and data-efficient pipeline for medical image analysis networks to enhance their accuracy and reliability in computer-assisted medical diagnoses.

Self-supervised learning, a technique for constructing feature embedding spaces by designing pretext tasks, has emerged as a promising solution for addressing the issue of label deficiency. One representative methodology for self-supervised learning is the masked autoencoder (MAE) [11]. MAEs learn to reconstruct input data after randomly masking certain input features. This approach has been successfully deployed in various applications, including image denoising, text completion, anomaly detection, and feature learning. In the field of medical image analysis, MAE pre-training has also been found to be effective [32]. Nevertheless, these studies have a limitation in that they require a large set of unlabeled data and do not prioritize improving output reliability, which may undermine their practicality in the real world.

In this paper, we propose **M**asked **M**ulti-view with **Swin** (SwinMM), the first comprehensive multi-view pipeline for self-supervised medical image segmentation. We draw inspiration from previous studies [26,28,31,33] and aim to enhance output reliability and data utilization by incorporating multi-view learning into the self-supervised learning pipeline. During the pre-training stage, the proposed approach randomly masks 3D medical images and creates various observations from different views. A masked multi-view encoder processes these observations simultaneously to accomplish four proxy tasks: image reconstruction, rotation, contrastive learning, and a novel proxy task that utilizes a mutual learning paradigm to maximize consistency between predictions from different views. This approach effectively leverages hidden multi-view information from 3D medical data and allows the encoder to learn enriched high-level representations of the original images, which benefits the downstream segmentation task. In the fine-tuning stage, different views from the same image are encoded into a series of representations, which will interact with each other in a specially designed cross-view attention block. A multi-view consistency loss is imposed to generate aligned output predictions from various perspectives, which enhances the reliability and precision of the final output. The complementary nature of the different views used in SwinMM results in higher precision, requiring less training data and annotations, which holds significant potential for advancing the state-of-the-art in this field. In summary, the contributions of our study are as follows:

- We present SwinMM, a unique and data-efficient pipeline for 3D medical image analysis, providing the first comprehensive multi-view, self-supervised approach in this field.
- Our design includes a masked multi-view encoder and a novel mutual learning-based proxy task, facilitating effective self-supervised pretraining.
- We incorporate a cross-view decoder for optimizing the utilization of multi-view information via a cross-attention block.
- SwinMM delivers superior performance with an average Dice score of 86.18% on the WORD dataset, outperforming other leading segmentation methods in both data efficiency and segmentation performance.

2 Method

Figure 1 provides an overview of SwinMM, comprising a masked multi-view encoder and a cross-view decoder. SwinMM creates multiple views by randomly masking an input image, subsequently feeding these masked views into the encoder for self-supervised pre-training. In the fine-tuning stage, we architect a cross-view attention module within the decoder. This design facilitates the effective utilization of multi-view information, enabling the generation of more precise segmentation predictions.

2.1 Pre-training

Masked Multi-view Encoder. Following [11], we divided the 3D images into sub-volumes of the same size and randomly masked a portion of them, as demonstrated in Fig. 2. These masked 3D patches, from different perspectives, were then utilized for self-supervised pretraining by the masked multi-view encoder. As shown in Fig. 1, the encoder is comprised of a patch partition layer, a patch embedding layer, and four Swin Transformer layers [17]. Notably, unlike typical transformer encoders, our masked multi-view encoder can process multiple inputs from diverse images with varying views, making it more robust for a broad range of applications.

Pre-training Strategy. To incorporate multiple perspectives of a 3D volume, we generated views from different observation angles, including axial, coronal, and sagittal. Furthermore, we applied rotation operations aligned with each perspective, consisting of angles of 0°, 90°, 180°, and 270° along the corresponding direction. To facilitate self-supervised pre-training, we devised four proxy tasks. The reconstruction and rotation tasks measure the model's performance on each input individually, while the contrastive and mutual learning tasks enable the model to integrate information across multiple views.

- **The reconstruction task** compares the difference between unmasked input $\mathcal{X}$ and the reconstructed image y^{rec}. Following [11], we adopt Mean-Square-Error (MSE) to compute the reconstruction loss:

$$\mathcal{L}_{rec} = (\mathcal{X} - y^{rec})^2. \tag{1}$$

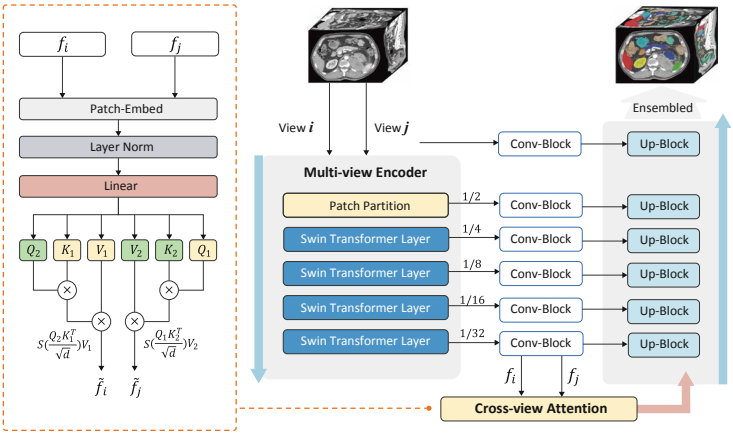

Fig. 1. Overview of our proposed SwinMM. The Conv-Blocks convolve the latent representations obtained from different levels of the masked multi-view encoder, adapting their feature size to match that of the corresponding decoder layer. The Up-Blocks perform deconvolution to upsample the feature maps.

– **The rotation task** aims to detect the rotation angle of the masked input along the axis of the selected perspective, with possible rotation angles of 0°, 90°, 180°, and 270°. The model's performance is evaluated using cross-entropy loss, as shown in Eq. 2, where y^{rot} and y_r represent the predicted probabilities of the rotation angle and the ground truth, respectively.

$$\mathcal{L}_{rot} = -\sum_{r=1}^{R} y_r \log(y^{rot}). \tag{2}$$

– **The contrastive learning task** aims to assess the effectiveness of a model in representing input data by comparing high-level features of multiple views. Our working assumption is that although the representations of the same sample may vary at the local level when viewed from different perspectives, they should be consistent at the global level. To compute the contrastive loss, we use cosine similarity $sim(\cdot)$, where y_i^{con} and y_j^{con} represent the contrastive pair, t is a temperature constant, and 1 is the indicator function.

$$\mathcal{L}_{con} = -\log \frac{\exp(sim(y_i^{con}, y_j^{con})/t)}{\sum_k^{2N} 1_{k \neq i} \exp(sim(y_i^{con}, y_k^{con})/t)}. \tag{3}$$

– **The mutual learning task** assesses the consistency of reconstruction results from different views to enable the model to learn aligned information from multi-view inputs. Reconstruction results are transformed into a uniform perspective and used to compute a mutual loss $\mathcal{L}_{mul}$, which, like the reconstruction task, employs the MSE loss. Here, $y^{rec}i$ and $y^{rec}j$ represent the predicted reconstruction from views i and j, respectively.

$$\mathcal{L}_{mul} = (y_i^{rec} - y_j^{rec})^2. \tag{4}$$

The total pre-training loss is as shown in Eq. 5. The weight coefficients α_1, α_2, α_3 and α_4 are set equal in our experiment ($\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$).

$$\mathcal{L}_{pre} = \alpha_1 \mathcal{L}_{rec} + \alpha_2 \mathcal{L}_{rot} + \alpha_3 \mathcal{L}_{con} + \alpha_4 \mathcal{L}_{mul}. \tag{5}$$

2.2 Fine-Tuning

Cross-View Decoder. The structure of the cross-view decoder, comprising Conv-Blocks for skip connection, Up-Blocks for up-sampling, and a Cross-view Attention block for views interaction, is depicted in Fig. 1. The Conv-Blocks operate on different layers, reshaping the latent representations from various levels of the masked multi-view encoder by performing the convolution, enabling them to conform to the feature size in corresponding decoder layers ($\frac{H}{2^i}, \frac{W}{2^i}, \frac{D}{2^i}, i = 0, 1, 2, 3, 4, 5$). At the bottom of the U-shaped structure, the cross-view attention module integrates the information from two views. The representations at this level are assumed to contain similar semantics. The details of the cross-view attention mechanism are presented in Fig. 1 and Eq. 6. In the equation, f_i and f_j denote the representations of different views, while Q_i, K_i, and V_i refer to the *query*, *key*, and *value* matrices of f_i, respectively.

$$\text{Cross Attention}(f_i, f_j) = [\text{Softmax}\left(\frac{Q_i K_j^\top}{\sqrt{d}}\right) V_j, \text{Softmax}\left(\frac{Q_j K_i^\top}{\sqrt{d}}\right) V_i]. \tag{6}$$

Multi-view Consistency Loss. We assume consistent segmentation results should be achieved across different views of the same volume. To quantify the consistency of the multi-view results, we introduce a consistency loss $\mathcal{L}_{mc}$, calculated using KL divergence in the fine-tuning stage, as in previous work on mutual learning [29]. The advantage of KL divergence is that it does not require class labels and has been shown to be more robust during the fine-tuning stage. We evaluate the effectiveness of different mutual loss functions in an ablation study (see supplementary). The KL divergence calculation is shown in Eq. 7:

$$\mathcal{L}_{MC} = D_{KL}(V_i \| V_j) = \sum_{m=1}^{N} V_i(x_m) \cdot \log \frac{V_i(x_m)}{V_j(x_m)}, \tag{7}$$

where $V_i(x_m)$ and $V_j(x_m)$ denote the different view prediction of m-th voxel. N represents the number of voxels of case x. $V_i(x)$ and $V_j(x)$ denote different view prediction of case x. We measure segmentation performance using $\mathcal{L}_{DiceCE}$, which combines Dice Loss and Cross Entropy Loss according to [24].

$$\mathcal{L}_{DiceCE} = 1 - \sum_{m=1}^{N} \left(\frac{2|p_m \cap y_m|}{N(|p_m| + |y_m|)} + \frac{y_m \log(p_m)}{N}\right), \tag{8}$$

where p_m and y_i respectively represent the predicted and ground truth labels for the m-th voxel, while N is the total number of voxels. We used $\mathcal{L}_{fin}$ during the fine-tuning stage, as specified in Eq. 9, and added weight coefficients β_{DiceCE} and β_{mc} for different loss functions, both set to a default value of 1.

$$\mathcal{L}_{fin} = \beta_{DiceCE} \mathcal{L}_{DiceCE} + \beta_{MC} \mathcal{L}_{MC}. \tag{9}$$

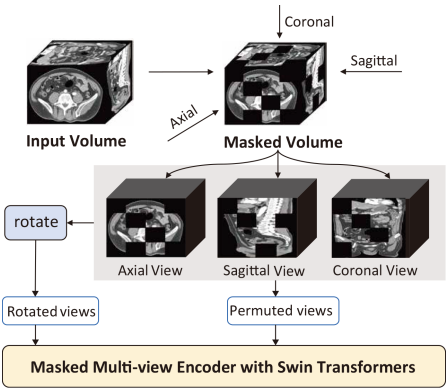

Fig. 2. SwinMM's pre-training stage.

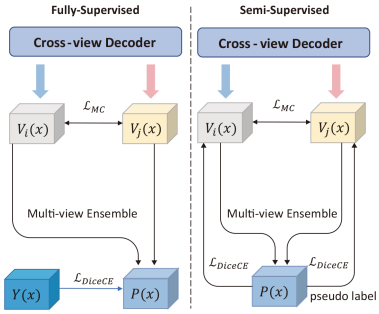

Fig. 3. The Fully-supervised/Semi-supervised pipeline with SwinMM.

2.3 Semi-supervised Learning with SwinMM

As mentioned earlier, the multi-view nature of SwinMM can substantially enhance the reliability and accuracy of its final output while minimizing the need for large, high-quality labeled medical datasets, making it a promising candidate for semi-supervised learning. In this study, we propose a simple variant of SwinMM to handle semi-supervision. As depicted in Fig. 3, we leverage the diverse predictions from different views for unlabeled data and generate aggregated pseudo-labels for the training process. Compared to single-view models, SwinMM's multi-view scheme can alleviate prediction uncertainty by incorporating more comprehensive information from different views, while ensemble operations can mitigate individual bias.

3 Experiments

Datasets and Evaluation. Our pre-training dataset includes 5833 volumes from 8 public datasets: AbdomenCT-1K [19], BTCV [13], MSD [1], TCIA-Covid19 [9], WORD [18], TCIA-Colon [14], LiDC [2], and HNSCC [8]. We choose two popular datasets, WORD (The Whole abdominal ORgan Dataset) and ACDC [3] (Automated Cardiac Diagnosis Challenge), to test the downstream segmentation performance. The accuracy of our segmentation results is evaluated using two commonly used metrics: the Dice coefficient and Hausdorff Distance (HD).

Implementation Details. Our SwinMM is trained on 8 A100 Nvidia GPUs with 80G gpu memory. In the pre-training process, we use a masking ratio of 50%, a batch size of 2 on each GPU, and an initial learning rate of 5e−4 and weight decay of 1e−1. In the finetuning process, we apply a learning rate of 3e−4

and a layer-wise learning rate decay of 0.75. We set 100K steps for pre-training and 2500 epochs for fine-tuning. We use the AdamW optimizer and the cosine learning rate scheduler in all experiments with a warm-up of 50 iterations to train our model. We follow the official data-splitting methods on both WORD and ACDC, and report the results on the test dataset. For inference on these datasets, we applied a double slicing window inference, where the window size is $64 \times 64 \times 64$ and the overlapping between windows is 70%.

3.1 Results

Comparing with SOTA Baselines. We compare the segmentation performance of SwinMM with several popular and prominent networks, comprising fully supervised networks, i.e., U-Net [22], Swin UNet [17], VT-UNet [21], UNETR [10], DeepLab V3+ [6], ESPNet [20], DMFNet [4], and LCOVNet [30], as well as self-supervised method Swin UNETR [24]. As shown in Table 1 and Table 2, our proposed SwinMM exhibits remarkable efficiency in medical segmentation by surpassing all other methods and achieves higher average Dice (86.18% on WORD and 90.80% on ACDC) and lower HD (9.35 on WORD and 6.37 on ACDC).

Single View vs. Multiple Views. To evaluate the effectiveness of our proposed multi-view self-supervised pretraining pipeline, we compared it with the state-of-the-art self-supervised learning method SwinUNETR [24] on WORD [18] dataset. Specifically, two SwinUNETR-based methods are compared: using fixed single views (Axial, Sagittal, and Coronal) and using ensembled predictions from multiple views (denoted as SwinUNETR-Fuse). Our results, presented in Table 3, show that our SwinMM surpasses all other methods including SwinUNETR-Fuse, highlighting the advantages of our unique multi-view designs. Moreover, by incorporating multi-view ensemble operations, SwinMM can effectively diminish the outliers in hard labels and produce more precise outputs, especially when dealing with harder cases such as smaller organs. The supplementary material provides qualitative comparisons of 2D/3D segmentation outcomes.

Table 1. Quantitative results of ACDC dataset. Note: RV - right ventricle, Myo - myocardium, LV - left ventricle.

Methods	DICE (%) ↑				HD ↓			
	RV	Myo	LV	Average	RV	Myo	LV	Average
U-Net [22]	54.17	43.92	60.23	52.77	24.15	35.32	60.16	39.88
Swin UNet [17]	78.50	77.92	86.31	80.91	11.42	5.61	7.42	8.12
VT-UNet [21]	80.44	80.71	89.53	83.56	11.09	5.24	6.32	7.55
UNETR [10]	84.52	84.36	92.57	87.15	12.14	5.19	4.55	7.29
Swin UNETR [24]	87.49	88.25	92.72	89.49	12.45	5.78	4.03	7.42
SwinMM	90.21	88.92	93.28	90.80	8.85	3.10	7.16	6.37

Table 2. Quantitative results of WORD dataset. Note: Liv - liver, Spl - spleen, Kid L - left kidney, Kid R - right kidney, Sto - stomach, Gal - gallbladder, Eso - esophagus, Pan - pancreas, Duo - duodenum, Col - colon, Int - intestine, Adr - adrenal, Rec - rectum, Bla - bladder, Fem L - left femur, Fem R - right femur.

Methods	Liv	Spl	Kid L	Kid R	Sto	Gal	Eso	Pan	Duo	Col	Int	Adr	Rec	Bla	Fem L	Fem R	**DICE (%)↑**	**HD↓**
UNETR [10]	94.67	92.85	91.49	91.72	85.56	65.08	67.71	74.79	57.56	74.62	80.4	60.76	74.06	85.42	89.47	90.17	79.77	17.34
CoTr [27]	95.58	94.9	93.26	93.63	89.99	76.4	74.37	81.02	63.58	84.14	86.39	69.06	80.0	89.27	91.03	91.87	84.66	12.83
DeepLab V3+ [6]	96.21	94.68	92.01	91.84	91.16	80.05	74.88	82.39	62.81	82.72	85.96	66.82	81.85	90.86	92.01	92.29	84.91	9.67
Swin UNETR [24]	96.08	95.32	94.20	94.00	90.32	74.86	76.57	82.60	65.37	84.56	87.37	66.84	79.66	92.05	86.40	83.31	84.34	14.24
ESPNet [20]	95.64	93.9	92.24	94.39	87.37	67.19	67.91	75.78	62.03	78.77	72.8	60.55	74.32	78.58	88.24	89.04	79.92	15.02
DMFNet [4]	95.96	94.64	94.7	94.96	89.88	79.84	74.1	81.66	66.66	83.51	86.95	66.73	79.26	88.18	91.99	92.55	85.1	7.52
LCOVNet [30]	95.89	95.4	95.17	95.78	90.86	78.87	74.55	82.59	68.23	84.22	87.19	69.82	79.99	88.18	92.48	93.23	85.82	9.11
SwinMM	96.30	95.46	93.83	94.47	91.43	80.08	76.59	83.60	67.38	86.42	88.58	69.12	80.48	90.56	92.16	92.40	86.18	9.35

Table 3. Quantitative results of WORD dataset. Abbreviations follows Table 2.

Methods	Liv	Spl	Kid L	Kid R	Sto	Gal	Eso	Pan	Duo	Col	Int	Adr	Rec	Bla	Fem L	Fem R	**DICE (%)↑**	**HD↓**
Swin UNETR Axi	96.08	95.32	94.20	94.00	90.32	74.86	76.57	82.60	65.37	84.56	87.37	66.84	79.66	92.05	86.40	83.31	84.34	14.24
Swin UNETR Sag	96.09	95.32	93.53	94.72	90.68	73.31	74.10	83.07	66.98	84.21	86.37	68.07	78.89	91.18	91.67	91.28	84.97	40.88
Swin UNETR Cor	96.12	95.49	93.91	94.80	90.25	71.78	75.27	82.83	66.26	84.07	86.98	66.23	79.38	90.93	88.09	86.74	84.32	14.02
Swin UNETR Fuse	96.25	95.71	94.20	94.85	91.05	74.80	77.04	83.73	67.36	85.15	87.69	67.84	80.29	92.31	90.44	89.36	85.50	13.87
SwinMM	96.30	95.46	93.83	94.47	91.43	80.08	76.59	83.60	67.38	86.42	88.58	69.12	80.48	90.56	92.16	92.40	86.18	9.35

Table 4. The ablation study of proxy tasks during pre-training.

Methods	Rec	Mut	Rot	Con	DICE (%) ↑	HD ↓
SwinMM (w/o pretraining)	–	–	–	–	84.78	11.77
SwinMM	✓	–	–	–	84.93	11.61
SwinMM	–	–	✓	✓	84.91	11.69
SwinMM	✓	✓	–	–	85.65	10.25
SwinMM	✓	–	✓	✓	85.19	10.98
SwinMM	✓	✓	✓	✓	85.74	9.56

Table 5. The ablation study of label ratios.

label ratio	Swin UNETR	SwinMM
10%	56.14	67.83
30%	70.65	78.91
50%	77.28	82.03
70%	81.07	83.25
90%	82.46	84.32
100%	83.13	84.78

3.2 Ablation Study

To fairly evaluate the benefits of our proposed multi-view design, we separately investigate its impact in the pre-training stage, the fine-tuning stage, as well as both stages. Additionally, we analyze the role of each pre-training loss functions.

Pre-training Loss Functions. The multi-view pre-training is implemented by proxy tasks. The role of each task can be revealed by taking off other loss functions. For cheaper computations, we only pre-train our model on 2639 volumes from 5 datasets (AbdomenCT-1K, BTCV, MSD, TCIA-Covid19, and WORD) in these experiments, and we applied a 50% overlapping window ratio, during testing time. As shown in Table 4, our proposed mutual loss brings a noticeable improvement in Dice (around 1%) over the original SwinUNETR setting. When combining all the proxy tasks, our SwinMM achieves the best performance.

Data Efficiency. The data efficiency is evaluated under various semi-supervised settings. Initially, a base model is trained from scratch with a proportion of supervised data from the WORD dataset for 100 epochs. Then, the base model finishes the remaining training procedure with unsupervised data. The proportion of supervised data (denoted as label ratio) varies from 10% to 100%. Table 5

shows SwinMM consistently achieves higher Dice (%) than SwinUNETR, and its superiority is more remarkable when training with fewer supervised data.

4 Conclusion

This paper introduces SwinMM, a self-supervised multi-view pipeline for medical image analysis. SwinMM integrates a masked multi-view encoder in the pre-training phase and a cross-view decoder in the fine-tuning phase, enabling seamless integration of multi-view information, thus boosting model accuracy and data efficiency. Notably, it introduces a new proxy task employing a mutual learning paradigm, extracting hidden multi-view information from 3D medical data. The approach achieves competitive segmentation performance and higher data-efficiency than existing methods and underscores the potential and efficacy of multi-view learning within the domain of self-supervised learning.

Acknowledgement. This work is partially supported by the Google Cloud Research Credits program.

References

1. Antonelli, M., et al.: The medical segmentation decathlon. Nat. Commun. **13**(1), 1–13 (2022)
2. Armato, S.G., III., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**, 2514–2525 (2018)
4. Chen, C., Liu, X., Ding, M., Zheng, J., Li, J.: 3D dilated multi-fiber network for real-time brain tumor segmentation in MRI. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 184–192. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_21
5. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. In: ICLR (2020)
8. Grossberg, A.J., et al.: Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Sci. Data **5**, 180173 (2018)
9. Harmon, S.A., et al.: Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat. Commun. **11**(1), 1–7 (2020)
10. Hatamizadeh, A., Yang, D., Roth, H.R., Xu, D.: UNETR: transformers for 3D medical image segmentation. In: WACV (2022)
11. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: CVPR (2022)

12. Hong, Q., et al.: A distance transformation deep forest framework with hybrid-feature fusion for CXR image classification. IEEE Trans. Neural Netw. Learn. Syst. (2023)
13. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: a survey. Med. Image Anal. **24**(1), 205–219 (2015)
14. Johnson, C.D., Chen, M., Toledano, A.Y., et al.: Accuracy of CT colonography for detection of large adenomas and cancers. Obstet. Gynecol. Surv. **64**, 35–37 (2009)
15. Kim, S., Nam, J., Ko, B.C.: ViT-NeT: interpretable vision transformers with neural tree decoder. In: ICML (2022)
16. Li, Z., Li, Y., Li, Q., et al.: LViT: language meets vision transformer in medical image segmentation. IEEE Trans. Med. Imaging (2023)
17. 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: ICCV (2021)
18. Luo, X., Liao, W., Xiao, J., et al.: WORD: a large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. Med. Image Anal. **82**, 102642 (2022)
19. Ma, J., Zhang, Y., Gu, S., et al.: AbdomenCT-1K: is abdominal organ segmentation a solved problem. IEEE Trans. Pattern Anal. Mach. Intell. (2021)
20. Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H.: ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 561–580. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_34
21. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A robust volumetric transformer for accurate 3D tumor segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13435. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_16
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
23. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med. Image Anal. **63**, 101693 (2020)
24. Tang, Y., et al.: Self-supervised pre-training of Swin transformers for 3D medical image analysis. In: CVPR (2022)
25. Wu, D., et al.: A learning based deformable template matching method for automatic rib centerline extraction and labeling in CT images. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 980–987. IEEE (2012)
26. Xia, Y., Yang, D., Yu, Z., et al.: Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med. Image Anal. **65**, 101766 (2020)
27. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 171–180. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_16
28. Zhai, P., Cong, H., Zhu, E., Zhao, G., Yu, Y., Li, J.: MVCNet: multiview contrastive network for unsupervised representation learning for 3-D CT lesions. IEEE Trans. Neural Netw. Learn. Syst. (2022)
29. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR (2018)

30. Zhao, Q., Wang, H., Wang, G.: LCOV-NET: a lightweight neural network for COVID-19 pneumonia lesion segmentation from 3D CT images. In: ISBI (2021)
31. Zhao, Z., et al.: MMGL: multi-scale multi-view global-local contrastive learning for semi-supervised cardiac image segmentation. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 401–405. IEEE (2022)
32. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image analysis. arXiv preprint arXiv:2203.05573 (2022)
33. Zhou, Y., et al.: Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In: WACV (2019)