FDAS: Foundation Model Distillation and Anatomic Structure-aware Multi-task Learning for Self-Supervised Medical Image Segmentation

Xiaoran Qi[1], Guoning Zhang[2], Jianghao Wu[3], Shaoting Zhang[3,4], Xiaorong Hou[1(✉)], and Guotai Wang[3,4(✉)]

[1] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China
houxr@uestc.edu.cn
[2] School of Information Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China
[3] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[4] Shanghai Artificial Intelligence Laboratory, Shanghai, China
guotai.wang@uestc.edu.cn

Abstract. Self-Supervised Learning (SSL) has shown promising results in medical image segmentation, offering advanced performance with minimal annotations. However, the absence of semantics during pre-training limits the performance of downstream tasks (e.g., organ segmentation). To address this issue, we propose a novel SSL framework via **F**oundation model **D**istillation and **A**natomic **S**tructure-aware multi-task learning (FDAS) for medical image segmentation. Specifically, we distill knowledge from the Segment Anything Model (SAM) and propose SAM-guided anatomic Structure-aware Masked Image Modeling (S2MIM), which randomly masks multiple anatomic structures in the image to enrich representation learning. For better pre-training, we introduce anatomic structure-aware multi-task learning, which integrates reconstruction and segmentation of anatomic structure-fused images to capture richer semantic information, along with fusion-based contrastive learning to preserve the semantic integrity and discriminative power of the learned representations. Experiments on two applications (cardiac MRI segmentation and fetal brain MRI segmentation) demonstrate that our method effectively improved the representation learning and outperformed several state-of-the-art SSL methods. The code is available at https://github.com/HiLab-git/FDAS.

Keywords: Self-Supervised Learning · Foundation Model · Image Segmentation.

1 Introduction

Deep learning has significantly advanced medical image segmentation, playing a vital role in computer-aided diagnosis [1, 16, 18, 23]. However, this success relies heavily on large amounts of expensive annotations for training [5, 15, 24].

Recent researches indicate that Self-Supervised Learning (SSL) is a promising paradigm for its superiority in learning representations without relying on large-scale annotations [3,6,13,20,25]. Contrastive Learning (CL) has been widely used in SSL [6, 7, 19]. SimCLR [6] learns representations by using contrastive loss to maximize agreement between augmented views of the same image while pushing apart views of different images. MoCo [7, 10] builds a dynamic dictionary with a queue and a momentum encoder, enabling consistent contrastive learning for SSL. VoCo [19] is a volume contrast framework that utilizes contextual position priors to learn consistent semantic representations for pre-training. Masked Image Modeling (MIM) is also a typical pretext task in SSL [4, 9, 21, 22]. MAE [9] and SimMIM [21] learn representations by masking parts of input images and reconstructing the missing pixels. SurgNet [4] introduces local semantic consistency to generate pseudo-masks and perform guided MIM, enhancing feature learning for pre-training. HybridMIM [22] proposes a two-level masking hierarchy for masked image modeling, facilitating semantic learning at multiple levels.

However, these methods often overlook target structure semantics during pre-training. For example, CL methods rely on simple augmentations like rotation, scaling, or global-local comparisons. MIM methods reconstruct masked regions using pre-defined shapes, which lack semantic relevance to the target, limiting SSL performance. Recently, the Segment Anything Model (SAM) [11] has been proposed as a foundation model for natural image segmentation, achieving impressive zero-shot segmentation ability. Can we leverage foundation model (e.g., SAM [11]) to further strengthen the representation learning for SSL?

Along this direction, we propose a novel SSL framework via **F**oundation model **D**istillation and **A**natomic **S**tructure-aware multi-task learning (FDAS) to further enhance representation learning for medical image segmentation. Our contributions are summarized as follows: (1) We propose SAM-guided anatomic Structure-aware Masked Image Modeling (S2MIM), which distills SAM semantic segmentation knowledge and leverages its zero-shot segmentation to randomly mask multiple anatomic structures in the input image, enhancing MIM with enriched target semantics. (2) Based on S2MIM, we introduce Image Fusion-driven Reconstruction (IFR) to reconstruct fused images and SAM Knowledge Distillation (SKD) to segment anatomic targets, which can capture richer semantics for representation learning in pre-training. (3) We also propose the Fusion-based Contrastive Learning (FCL) strategy to maximize agreement between anatomic structure-fused images and the corresponding original images while pushing apart views of different images, which reserves the semantic integrity and the discriminativeness of the learned representations. This strategy can effectively improve the capabilities of learning anatomic structural-semantic.

Extensive experiments on two applications (cardiac MRI segmentation and fetal brain MRI segmentation) showed that our method can effectively improve the representation learning in SSL. It outperformed several state-of-the-art SSL methods for medical image segmentation.

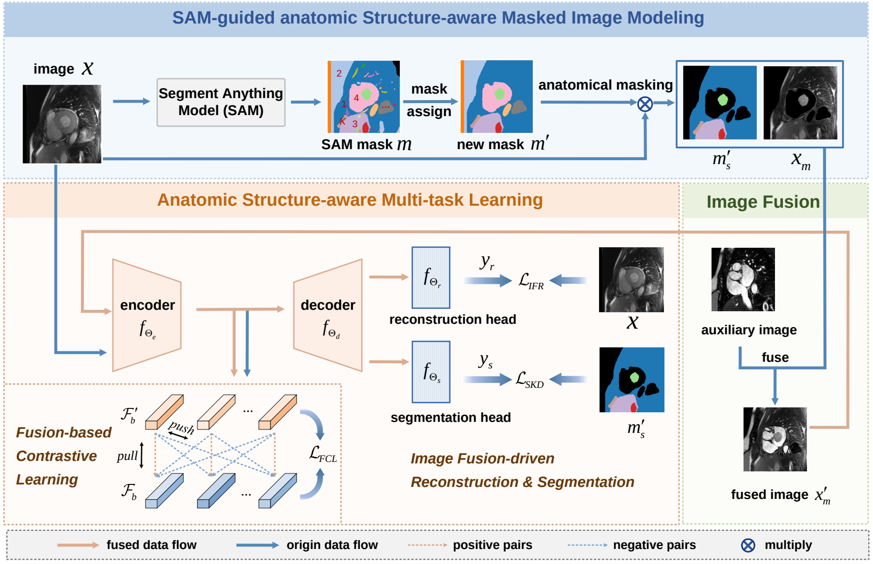

Fig. 1. Overview of our FDAS. We use SAM-guided anatomic Structure-aware Masked Image Modeling (S2MIM) to generate masked image x_m through randomly anatomic structural masking in the input x. The fused image x'_m can be used for anatomic structure-aware multi-task learning, which includes reconstruction loss $\mathcal{L}_{IFR}$, segmentation loss $\mathcal{L}_{SKD}$ and contrastive loss $\mathcal{L}_{FCL}$.

2 Method

Fig. 1 shows the overview of our FDAS. We distill SAM knowledge to guide anatomic structure-aware masked image modeling, which randomly masks multiple anatomic structures in the input image. Through image fusion-driven reconstruction and segmentation, the fused images are reconstructed and segmented to enhance anatomic semantics in representation learning. Fusion-based contrastive learning aligns fused images with their original images while separating different images.

2.1 SAM-guided anatomic Structure-aware Masked Image Modeling

Existing SSL methods based on MIM demonstrate limited performance due to the lack of target semantics in the masking process. Considering this issue, we propose SAM-guided anatomic Structure-aware Masked Image Modeling (S2MIM) to achieve better representation learning. It distills the knowledge from the powerful SAM [11] and provides valuable anatomic semantics for a lightweight model (e.g., UNet [12]).

Specifically, for an input image $x \in \mathbb{R}^{H \times W}$, where H and W are the height and width, respectively. We use SAM's automatic mask generator to produce masks $m \in \{0,1\}^{M\times H\times W}$ for x, with each sub-mask denoted as $m_i \in \{0,1\}^{H\times W}$ ($i = 1, 2, \ldots, M$), where M denotes the total number of sub-masks. To enhance the consistency of m and reduce noise, we apply a preprocessing step. Motivated by the 95% variance contribution principle from Principal Component Analysis (PCA) [8], we first sort the sub-masks m_i by area and then determine the minimal number k required to cover at least 95% of the total area, in order to mitigate the influence of small noisy areas in m. We average k across all images and rounding up before taking the nearest integer to obtain K of principal representation sub-masks, which is formally defined as:

$$K = \left\lceil \frac{1}{N}\sum_{n=1}^{N}\left(\min_{k}\Big(\frac{\sum_{i=1}^{k}\sum_{p=1}^{H\times W} m_{i,n}^{*p}}{\sum_{i=1}^{M}\sum_{p=1}^{H\times W} m_{i,n}^{*p}} \geq \alpha\Big)\right)\right\rceil, \tag{1}$$

where N is the number of pre-training images, m_i^* is the sorted sub-masks, $\alpha = 95\%$ and $\lceil\cdot\rceil$ denotes rounding up to the nearest integer. We remain the top K largest sub-masks in each m that are referred to as anchor sub-masks. The remaining sub-masks in m are assigned to the nearest anchor sub-mask based on their Euclidean distance. As a result, we obtain a new mask $m' \in \{0,1\}^{K\times H\times W}$. If $M < K$, the extra $(K - M)$ channels are padded with zeros.

We randomly select $r \in (0,1)$ of the sub-masks in m' and mask them to obtain the anatomic structure mask map m'_s. Formally, we define m'_s as:

$$m'_s = m' \odot \mathbf{M}, \tag{2}$$

where $\odot$ denotes element-wise multiplication at the pixel level. $\mathbf{M} \in \{0,1\}^K$ is a selection mask, with r of sub-masks in m' randomly set to 0, and the remaining sub-masks are set to 1.

Then, we obtain the anatomic structure-masked image x_m by applying m'_s to the original image x. This is formally defined as:

$$x_m = x \odot \big(\sum_{k=1}^{K} {m'_s}^{k}\big). \tag{3}$$

2.2 Image Fusion-driven Reconstruction

The quality of feature representations learned during the pre-training phase is essential for SSL. To improve this, we propose multi-task learning that integrates three complementary tasks: Image Fusion-driven Reconstruction (IFR), SAM Knowledge Distillation (SKD) for segmentation, and Fusion-based Contrastive Learning (FCL). We employ a general encoder (f_{Θ_e}), a general decoder (f_{Θ_d}), a lightweight reconstruction head (f_{Θ_r}) and a simple segmentation head (f_{Θ_s}) to achieve multi-task learning in SSL.

Specifically, for IFR, we build upon our S2MIM and propose an image fusion-driven reconstruction to recover the anatomic structural semantics of different images, which reduces the variance of the masked image x_m and enhances the robustness of the reconstruction process. Let x_j be an auxiliary image randomly sampled from other images within the same batch as x_m. For image fusion, we use x_j to fill the masked regions in x_m. For each pixel location p, the anatomic structure-fused image $x'_m(p)$ is defined as follows:

$$x'_m(p) = \begin{cases} x_j(p), & \text{if } x_m(p) = 0, \\ x_m(p), & \text{otherwise.} \end{cases} \tag{4}$$

The reconstruction output y_r of x'_m is obtained by:

$$y_r = f\big(x'_m; \Theta_e, \Theta_d, \Theta_r\big). \tag{5}$$

Formally, the image fusion-driven reconstruction loss is given by:

$$\mathcal{L}_{IFR} = \frac{1}{H \times W} \sum_{p=1}^{H \times W} \big(y_r(p) - x(p)\big)^2. \tag{6}$$

2.3 SAM Knowledge Distillation for Segmentation

To capture richer anatomic structural information in SSL, we introduce SAM Knowledge Distillation (SKD) for segmentation, which distills the powerful prior knowledge of SAM and leverages its superiority in zero-shot segmentation to guide the learning process. For a given anatomic structure-fused input image x'_m, the segmentation output is defined as:

$$y_s = f\big(x'_m; \Theta_e, \Theta_d, \Theta_s\big). \tag{7}$$

We use m'_s as the label to supervise y_s and employ the Dice loss for supervision, which is defined as:

$$\mathcal{L}_{SKD} = 1 - \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{p=1}^{H \times W} 2 \cdot \big(y_s^k(p)\big) \cdot \big({m'_s}^k(p)\big)}{\sum_{p=1}^{H \times W} \big(y_s^k(p) + {m'_s}^k(p)\big) + \epsilon}, \tag{8}$$

where each of K one-hot regions is taken as a foreground class, sorted by region size, and $\epsilon = 10^{-7}$ is a small number for numerical stability.

2.4 Fusion-based Contrastive Learning

To preserve the semantic integrity and the global feature representation ability, we propose Fusion-based Contrastive Learning (FCL) for better pre-training.

Let the batch size be B, with a batch of original images $X = \{x_b\}_{b=1}^{B}$ and their corresponding fused images $X' = \{x'_{m,b}\}_{b=1}^{B}$. For each sample $x'_{m,b}$ and x_b, the feature embeddings obtained by f_{Θ_e} are denoted as $\mathcal{F}'_b$ and $\mathcal{F}_b$, respectively. For each sample b in the batch, the positive pair is $(\mathcal{F}'_b, \mathcal{F}_b)$, and the negative pairs are $(\mathcal{F}'_b, \mathcal{F}_g)$ for all $g \neq b$. Then, the Fusion-based Contrastive Loss based on the InfoNCE [10] loss is formulated as:

$$\mathcal{L}_{FCL} = -\frac{1}{B} \sum_{b=1}^{B} \log \frac{\exp\big(sim(\mathcal{F}'_b, \mathcal{F}_b)/\tau\big)}{\exp\big(sim(\mathcal{F}'_b, \mathcal{F}_b)/\tau\big) + \sum_{g=1, g \neq b}^{B} \exp\big(sim(\mathcal{F}'_b, \mathcal{F}_g)/\tau\big)}, \tag{9}$$

where $sim(\cdot, \cdot)$ denotes the cosine similarity function, and τ is the temperature coefficient, which is set to 0.1. The overall loss of our FDAS for pre-training is:

$$\mathcal{L} = \mathcal{L}_{IFR} + \mathcal{L}_{SKD} + \lambda \mathcal{L}_{FCL}, \tag{10}$$

where λ is the balancing hyper-parameter.

3 Experiment

3.1 Experimental Details

Dataset and Metrics Our experiments used the public Multi-Centre, Multi-Vendor, and Multi-Disease Cardiac Image Segmentation dataset (M&MS [2]) and a private Fetal Brain (FB) dataset [17]. The target tissues of the M&MS dataset are the Left Ventricle (LV), Myocardium (MYO), and Right Ventricle (RV). For the FB dataset, the target tissue is the fetal brain. The M&MS dataset contains 694 MRI volumes and the FB dataset includes 112 volumes. We randomly divided the datasets into 70% for training, 10% for validation, and 20% for testing, discarding the annotations of the training set during pre-training. For downstream tasks, we randomly select a small portion of images with annotations from the training set. The performance was quantitatively measured by Dice score and Average Symmetric Surface Distance (ASSD) in 3D space.

Implementation Details Our FDAS followed the two-step SSL paradigm, i.e., self-supervised representation learning and fully-supervised downstream fine-tuning. We adopted classic 2D UNet [12] to demonstrate the effectiveness of our method. The image intensity was linearly normalized to [-1, 1] and each slice was resized to 256×256. During pre-training, the M&MS dataset was trained for 30k iterations and the FB dataset was trained for 60k iterations. We adopted the Adam optimizer with a momentum of 0.9 and an initial learning rate of 10^{-3}. All parameters Θ were optimized jointly. The hyper-parameters r and λ were optimized using the validation set, and all ablation studies were conducted with 10% of the training data for fine-tuning. Specifically, we selected $r = 70\%$ and $\lambda = 0.5$ for both datasets. All experiments were implemented with PyTorch, using an NVIDIA GeForce RTX 3090 GPU. The pre-trained model is deployed on the SenseCare platform [14] for clinical research.

3.2 Results

Comparison with Other Methods Our FDAS was compared with five state-of-the-art SSL methods: 1) **MoCo** [10] that uses momentum contrastive learning for self-supervised learning. 2) **SimCLR** [6] that leverages contrastive learning based on data augmentation. 3) **MAE** [9] that employs masked image modeling for SSL. 4) **HybridMIM** [22] that employs two level masking for masked image modeling. 5) **VoCo** [19] that utilizes contextual position priors to capture consistent semantic representations. We also compared our method with:

Table 1. Quantitative comparison of different methods using 10% data for fine-tuning (excluding "Upper Bound") on both datasets. † indicates a significant improvement (p-value ≤ 0.05 in paired t-test) from the best values obtained by existing methods.

Method	Dice ↑ (%)				ASSD ↓ (pixel)	
	M&MS(LV)	M&MS(MYO)	M&MS(RV)	FB	M&MS(Avg)	FB
From Scratch	76.54±20.12	64.83±18.94	71.39±22.60	74.11±12.23	1.56±2.62	4.19±4.65
SAM-FT [11]	74.50±12.90	65.50±10.89	83.04±7.42	80.57±8.79	1.03±0.56	2.20±2.46
MoCo [10]	82.39±14.93	73.41±13.65	79.80±15.68	82.74±11.48	0.97±1.32	2.92±5.11
SimCLR [6]	83.05±15.38	74.38±12.93	77.86±17.11	81.43±9.55	1.17±1.51	3.12±4.67
MAE [9]	85.01±11.57	74.33±12.39	81.05±17.04	89.94±6.54	0.68±0.67	1.08±0.89
HybridMIM [22]	84.95±11.73	75.85±10.38	77.44±14.29	84.80±12.42	1.10±2.02	4.13±10.15
VoCo [19]	86.28±4.64	75.63±4.85	79.64±5.48	84.38±5.59	0.76±0.27	1.65±1.07
Ours	**89.62±6.66**†	**80.16±9.51**†	**84.83±10.67**†	**92.55±4.14**†	**0.58±0.69**†	**0.86±1.11**
Upper Bound	89.47±6.85	80.53±8.31	85.76±11.37	94.85±2.25	0.44±0.34	0.89±1.90

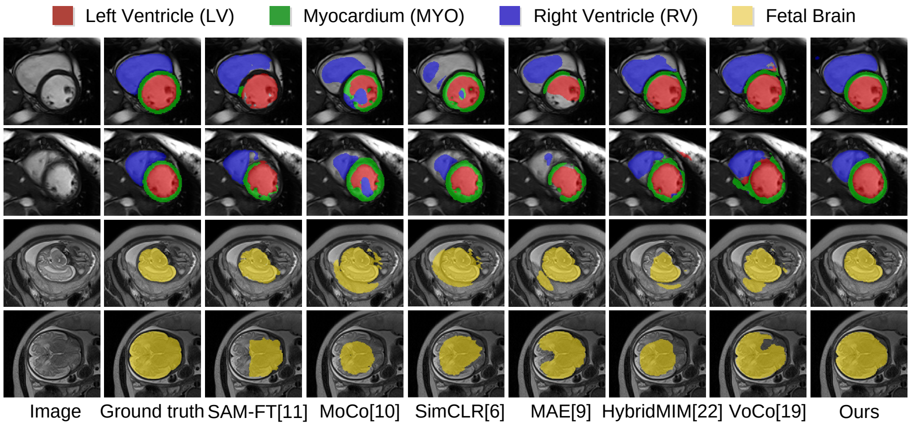

Fig. 2. Qualitative segmentation results of different methods. The first two rows and the last two rows are from the M&MS and FB datasets, respectively.

1) **From Scratch** that trains a randomly initialized network using 10% data for fully supervised learning. 2) **SAM-FT** [11] that uses annotated images to fine-tune SAM. 3) **Upper Bound** that uses 100% of the data to train the model. The results of downstream experiments on both datasets are shown in Table 1 and Fig. 3(a). Except for "Upper Bound", all of the methods use 10% of the data for fine-tuning. The existing methods only achieved moderate improvements compared to "From Scratch". In contrast, our method achieved significant improvements, outperforming most SSL methods in terms of Dice and ASSD. For example, compared to "From Scratch" in M&MS(LV) and FB segmentation, our method achieved average Dice scores of 89.62% and 92.55%, reflecting improvements of 13.08 and 18.44 percentage points, respectively. Fig. 3(a) shows the average Dice across LV, MYO, and RV on the M&MS dataset, our method outperformed existing methods across different data ratios, with the 20% data ratio result (85.97%) even surpassing "Upper Bound (85.25%)". Fig. 2 shows a

Table 2. Effectiveness of components in our FDAS. Baseline trains the model from scratch. S2MIM: Using SAM-guided anatomic structure-aware MIM for SSL. $\mathcal{L}_{IFR}$ and $\mathcal{L}_{SKD}$: Introducing reconstruction loss and segmentation loss for pre-training, respectively. $\mathcal{L}_{FCL}$: Using fusion-based contrastive loss for representation learning.

S2MIM	$\mathcal{L}_{IFR}$	$\mathcal{L}_{SKD}$	$\mathcal{L}_{FCL}$	Dice ↑ (%)				ASSD ↓ (pixel)
				LV	MYO	RV	Average	Average
baseline				76.54±20.12	64.83±18.94	71.39±22.60	70.92±20.55	1.56±2.62
✓				84.54±15.04	75.30±12.08	73.05±24.39	77.63±17.17	1.08±1.97
✓	✓			85.27±13.58	75.04±11.80	77.06±20.04	79.13±15.14	1.04±1.67
✓	✓	✓		86.53±12.08	78.00±12.20	80.73±15.65	81.75±13.31	0.65±1.08
✓	✓	✓	✓	**89.58±7.94**	**79.64±8.40**	**82.61±14.40**	**83.94±10.25**	0.66±0.93

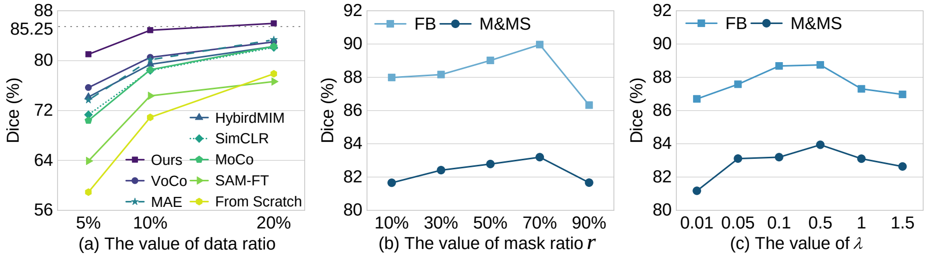

Fig. 3. (a) shows the average Dice scores across all target tissues on the M&MS dataset with varying data ratios. (b) and (c) show the effect of different mask ratios r and weights λ on the validation set of both datasets, respectively.

visual comparison between different SSL methods fine-tuned with 10% of the data. It can be observed that our segmentation results were closer to the ground truth and more complete.

Ablation Study There are two key hyper-parameters specific to our method: the mask ratio r and loss weight λ. We first investigated the effect of different r and the performance is shown in Fig. 3(b). We can observe that $r = 70\%$ achieved the best performance. Fig. 3(c) shows the performance with different λ values and the best λ was 0.5. To evaluate the effectiveness of components in our FDAS, we further investigated the impact of S2MIM, $\mathcal{L}_{IFR}$, $\mathcal{L}_{SKD}$, and $\mathcal{L}_{FCL}$ on the M&MS dataset. As shown in Table 2, each component of our method contributed to a performance improvement. The baseline achieved an average Dice of 70.92%. Only using S2MIM for SSL obtained the Dice of 77.63%, and additionally using $\mathcal{L}_{IFR}$ and $\mathcal{L}_{SKD}$ improved it to 79.13% and 81.75%, respectively. Our proposed method combining all these components achieved the highest Dice of 83.94%.

4 Conclusion

In this paper, we propose a novel SSL framework, FDAS, which leverages foundation model distillation and anatomic structure-aware multi-task learning to overcome the limitations of existing methods in medical image segmentation. We propose SAM-guided anatomic structure-aware masked image modeling to distill SAM knowledge and randomly mask multiple structures in the image to enhance anatomic semantics. Our method leverages anatomic structure-aware multi-task learning for better pre-training, integrating reconstruction and segmentation of fused images alongside fusion-based contrastive learning. Experiments on the public M&MS dataset and a private fetal brain dataset demonstrate the effectiveness of our method, achieving superior performance compared to several state-of-the-art SSL methods. In the future, it is of interest to apply our method to other downstream tasks.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No. 62271115) and the Natural Science Foundation of Sichuan Province (No. 2025ZNSFSC0455).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azad, R., et al.: Medical image segmentation review: The success of U-Net. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 10076–10095 (2024)
2. Campello, V.M., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging **40**(12), 3543–3554 (2021)
3. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. In: NeurIPS. vol. 33, pp. 12546–12558 (2020)
4. Chen, J., Li, M., Han, H., Zhao, Z., Chen, X.: SurgNet: Self-supervised pretraining with semantic consistency for vessel and instrument segmentation in surgical images. IEEE Trans. Med. Imaging **43**(4) (2024)
5. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. Med. Image Anal. **58**, 101539 (2019)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
7. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
8. Greenacre, M., et al.: Principal component analysis. Nat. Rev. Methods Prim. **2**(1), 100 (2022)
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
11. Kirillov, A., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
12. 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

13. Taleb, A., et al.: 3D self-supervised methods for medical imaging. In: NeurIPS. vol. 33, pp. 18158–18172 (2020)
14. Wang, G., Duan, Q., Shen, T., Zhang, S.: SenseCare: a research platform for medical image informatics and interactive 3D visualization. Frontiers in Radiology **4**, 1460889 (2024)
15. Wang, G., Wu, J., Luo, X., Liu, X., Li, K., Zhang, S.: MIS-FM: 3D medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. arXiv preprint arXiv:2306.16925 (2023)
16. Wu, J., Zhang, G., Qi, X., Wang, H., Liu, X., Wang, G.: Unsupervised domain adaptation for abdominal organ segmentation using pseudo labels and organ attention CycleGAN. In: MICCAI 2024 FLARE Challenge (2025)
17. Wu, J., et al.: UPL-SFDA: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation. IEEE Trans. Med. Imaging **42**(12), 3932–3943 (2023)
18. Wu, J., et al.: SAM-aware test-time adaptation for universal medical image segmentation. arXiv preprint arXiv:2506.05221 (2025)
19. Wu, L., Zhuang, J., Chen, H.: VoCo: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. In: CVPR. pp. 22873–22882 (2024)
20. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: UniMiSS+: Universal medical self-supervised learning from cross-dimensional unpaired data. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 10021–10035 (2024)
21. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: SimMIM: A simple framework for masked image modeling. In: CVPR. pp. 9653–9663 (2022)
22. Xing, Z., Zhu, L., Yu, L., Xing, Z., Wan, L.: Hybrid masked image modeling for 3d medical image segmentation. IEEE J. Biomed. Health Inform. **28**(4), 2115–2125 (2024)
23. Zhang, G., Qi, X., Yan, B., Wang, G.: IPLC: iterative pseudo label correction guided by SAM for source-free domain adaptation in medical image segmentation. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15011, pp. 351–360. Springer, Cham. https://doi.org/10.1007/978-3-031-72120-5_33
24. Zhou, H.Y., Lu, C., Chen, C., Yang, S., Yu, Y.: A unified visual information preservation framework for self-supervised pre-training in medical image analysis. IEEE Trans. Pattern Anal. Mach. Intell. **45**(7), 8020–8035 (2023)
25. Zhou, H.Y., Lu, C., Yang, S., Han, X., Yu, Y.: Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In: ICCV. pp. 3499–3509 (2021)