Fusing Dual Encoders: Single-source Domain Generalization with Extremely Few Annotations

Ruofan Wang[1], Jintao Guo[1], Jian Zhang[2,1], Lei Qi[3], and Yinghuan Shi[1,⋆]

[1] State Key Laboratory of Novel Software Technology,
National Institute of Healthcare Data Science, Nanjing University, China
rfwwang@gmail.com, syh@nju.edu.cn
[2] School of Intelligence Science and Technology, Nanjing University, China
[3] School of Computer Science and Engineering,
Key Lab of Computer Network and Information Integration (Ministry of Education),
Southeast University, China

Abstract. Considering the commonly existing domain shifts and label scarcity, single-source domain generalization (SDG) is a crucial and promising topic in medical image segmentation. SDG trains the model on one source domain and aims for generalization on the unseen target domain. However, previous methods rely on the quantity of training samples and perform poorly when only a few labeled training volumes are available, limiting the effective applicability in clinical practice. Thus, we concentrate on the challenging SDG setting with extremely few annotated samples and propose a **Me**dical **Du**al-encoder framework (MEDU). A dual-encoder U-shaped network incorporates two different encoders and fuses features via simple yet effective layers for learning representative features. We integrate pretrained SAM2 encoder with semantic knowledge for a proper initialization and resisting overfitting, proving effective in training with limited supervision. Furthermore, we introduce a perturbation consistency training strategy with perturbation operations and hierarchical consistency to learn domain-invariant features and alleviate discrepancies between training and inference. MEDU exceeds existing advanced methods in three challenging cross-domain settings concerning SDG with extremely few annotations. For example, on Abdominal MRI-CT, MEDU attains a Dice score of 81.75% with only three labeled training volumes, achieving an improvement of 12.60%. Our source code is available at https://github.com/wrf-nj/MEDU.

Keywords: Domain Generalization · Medical Image Segmentation · Extremely Few Annotations.

1 Introduction

Medical image segmentation [1,2,3], which identifies regions such as organs and tumors, is of great significance in medical image analysis. Recently, data-driven

⋆ Corresponding author.

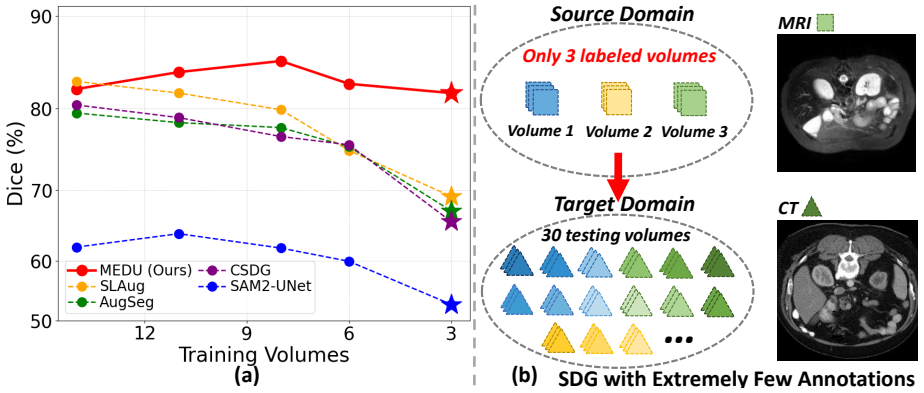

Fig. 1: Motivation of our proposed MEDU. (a) On Abdominal MRI-CT, previous methods rely on the number of labeled training volumes, while MEDU obtains a good performance even with extremely few labeled training volumes. (b) SDG with extremely few annotations (*e.g.,* the model is trained by only 3 labeled volumes on Abdominal MRI and tested on Abdominal CT).

deep learning methods have witnessed substantial progress [4], yet they encounter intrinsic issues concerning distribution shifts [5] and label scarcity. Specifically, inconsistencies (*e.g.,* imaging modalities) in the image acquisition process often result in domain distribution shifts [6,7]. In addition, pixel-wise labels for medical images are extremely scarce as manual annotation requires time investment and expertise [3]. These two problems hinder the effective implementation of medical image segmentation in clinical practice.

Previous works attempt to address distribution shifts via domain generalization (DG) [5,8]. Conventional DG [9,10] trains the model on multiple source domains and tests it on the unseen target domain, namely multi-source DG. To further ease label scarcity and privacy issues caused by involving multiple sources, single-source DG (SDG) [6,7,11] trains the model using one source domain and aims for generalization in unseen target domain(s). However, medical SDG still requires a considerable amount of training data (14, 21, 35 3D volumes for Abdominal MRI-CT, Abdominal CT-MRI, Cardiac bSSFP-LGE respectively) to obtain competitive performances [7]. Given the difficulty of annotation, ensuring generalization with scarce annotations remains a crucial topic.

We investigate the generalization ability of existing methods [6,7,12,13] in SDG with few annotations. As shown in Fig. 1 (a), we progressively reduce the number of training volumes by 20% per step and observe a significant tendency of performance degradation for previous methods, particularly when only a few labeled samples (*i.e.,* 3 volumes, accounting for 20%) are available on Abdominal MRI-CT [14,15]. *This issue arises from overfitting to the source domain with limited labeled training data and insufficient learning of domain-invariant semantic features.* Notably, as the knowledge of pretrained SAM2 helps resist overfitting, SAM2-UNet [13] performs better with fewer samples (100% vs. 80%).

When the training volumes are extremely limited, SAM2-UNet still gains poor performance. As achieving generalization with limited labeled data remains a challenge, we concentrate on SDG with extremely few annotations (Fig. 1(b)).

In medical image segmentation, SDG with extremely few annotations confronts two main challenges: limited annotated samples hamper the effectiveness of training and domain shifts further amplify the difficulty of generalization. For segmentation, many medical DG methods [6,7,9] build backbones based on U-Net [16], which utilizes multiscale features to capture semantics comprehensively. However, U-Net is not effective enough when trained on limited labeled samples and struggles to model long-range dependencies due to the locality of convolutions [17,18]. A natural solution is to integrate visual foundation models, which could serve for a proper initialization and help to reduce overfitting. SAM and SAM2 [19,20] are visual foundation models pre-trained on large-scale datasets and built upon Transformer-based architecture with advantages in modeling global relationships. Although exhibiting strong performances in downstream tasks [13,21,22], SAM and SAM2's potential in medical DG remains largely unexplored. Thus, we unify SAM2 and U-Net for effective training with limited supervision and design a training strategy to overcome domain shifts.

We propose a **Me**dical **Du**al-encoder framework (MEDU), which incorporates a dual-encoder U-shaped network for training with extremely few annotations and a perturbation consistency training strategy for learning domain-invariant features. A dual-encoder U-shaped network employs two distinct encoders (CNN-based encoder and Transformer-based encoder utilizing SAM2 [13]) and feature fusion modules with simple yet effective layers, integrating the capability of pretrained SAM2 for a proper initialization and resisting overfitting. Besides, a perturbation consistency training strategy is introduced with perturbation operations and hierarchical consistency for generalization. Perturbation operations utilize intensity transformations [12] and dropout to alleviate overfitting. Hierarchical consistency is employed on predictions acquired from multiscale features, which encourages the model to learn domain-invariant features and mitigates the differences between training and inference [23].

In this paper, three main contributions are listed as follows.

- A medical dual-encoder framework unifying both SAM2 and U-Net is proposed to solve single-source domain generalization with extremely few annotations in medical image segmentation.
- A novel dual-encoder U-shaped network with two encoders (Transformer-based and CNN-based) and feature fusion modules integrates features to learn semantic representations with few training samples.
- A perturbation consistency training strategy with perturbation operations and hierarchical consistency encourages the model to capture domain-invariant features and mitigates discrepancies between training and inference.

Extensive experiments conducted on three cross-domain settings demonstrate the effectiveness of our proposed MEDU. For example, on Abdominal MRI-CT, MEDU achieves a Dice score of 81.75% when trained with only three labeled volumes, yielding an improvement of 12.60% over the previous advanced method.

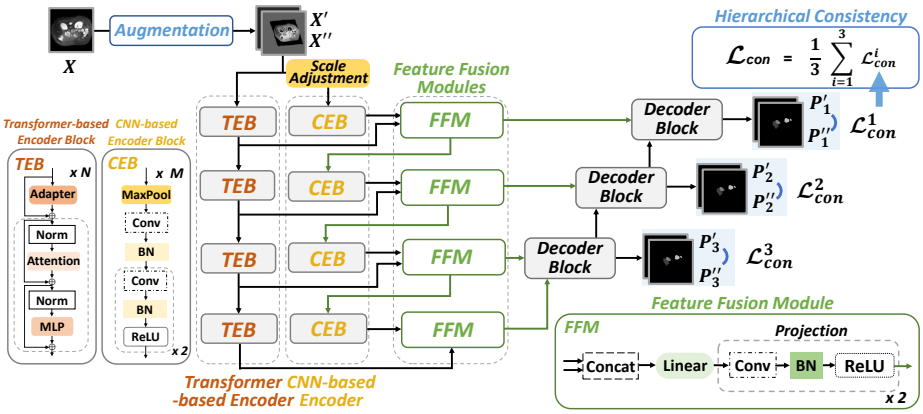

Fig. 2: Overview of our proposed MEDU, employing a dual-encoder U-shaped network and a perturbation consistency training strategy.

2 Method

SDG with extremely few annotations trains the model with limited labeled samples on one source domain and aims for the generalization in the unseen target domain. As illustrated in Fig. 2, we propose a medical dual-encoder framework (MEDU) with a dual-encoder U-shaped network and a perturbation consistency training strategy for SDG with extremely few annotations.

2.1 A Dual-encoder U-shaped Network

A dual-encoder U-shaped network is proposed, inspired by [19,13,16], to unify SAM2 and U-Net for enhanced representational capability with extremely few annotations. It mainly comprises scale adjustment, two encoders, feature fusion modules (FFMs), and one decoder. Transformer-based encoder and CNN-based encoder each consist of four layers of blocks, referred to as TEB and CEB respectively. Transformer-based encoder utilizes frozen Hiera blocks pre-trained by SAM2 and inserted adapters for fine-tuning [13], which facilitates model initialization and overfitting mitigation. CNN-based encoder leverages convolutions [16,9] for learning and follows scale adjustment. After encoding the inputs, four FFMs, which share similar structures but operate at different scales, are utilized to fuse features from the above two encoders. The decoder consists of three layers, each containing receptive field blocks and Convolution-Batch Normalization-ReLU combinations, while simultaneously generating predictions [13].

Scale Adjustment. To align input scales of CNN-based encoder with those of Transformer-based encoder for subsequent feature fusion, scale adjustment is introduced. It also functions as encoding layers and mainly utilizes convolutions, comprising a CNN-based encoder block (CEB) without max-pooling, followed by another CEB. As depicted in Fig. 2, CEB utilizes Max-pooling, Convolution, Batch Normalization, and ReLU, following U-Net [16,9].

Feature Fusion Module (FFM). Four FFMs with similar structures are employed to fuse multiscale outputs from the two different encoders. For the i-th ($i \in \{1, 2, 3, 4\}$) layer, $\mathcal{F}_t^i$ and $\mathcal{F}_c^i \in \mathbb{R}^{B\times C\times H\times W}$ denote the output features of Transformer-based encoder and CNN-based encoder respectively. We first concatenate $\mathcal{F}_t^i$ and $\mathcal{F}_c^i$ in the second dimension to produce the hybrid feature $\mathcal{F}_h^i \in \mathbb{R}^{B\times 2C\times H\times W}$. A learnable linear function $Linear(\cdot) : \mathbb{R}^{B\times 2C\times H\times W} \rightarrow \mathbb{R}^{B\times C\times H\times W}$ is employed to weight $\mathcal{F}_h^i$, perform dimension reduction, and generate $\mathcal{F}_h^{i'}$. Subsequently, $\mathcal{F}_h^{i'}$ is passed through a projection $P(\cdot) : \mathbb{R}^{B\times C\times H\times W} \rightarrow \mathbb{R}^{B\times C\times H\times W}$ with two layers, each consisting of Convolution, Batch Normalization, and ReLU [24]. $P(\cdot)$ preserves feature size while assisting in learning semantic representations. FFMs function as a component for effective encoding as well as learning representative semantic features.

The proposed dual-encoder U-shaped network fuses features from two distinct encoders to learn semantic representations, leveraging the pretrained SAM2 for a good initialization and easing overfitting in the source domain. Thus, it exhibits effectiveness when trained with few annotated samples.

2.2 A Perturbation Consistency Training Strategy

A perturbation consistency training strategy is introduced for generalization, which utilizes *Perturbation Operations* to produce variations and *Hierarchical Consistency* to enforce consistency in predictions of the same input under varying transformations and alleviate discrepancies between training and inference.

Perturbation Operations. Intensity perturbation is applied on images by randomly stacking multiple transformations [12], which introduces diverse styles. For the batch of input X, we utilize intensity perturbation twice to produce augmented batches $X^{'}$ and $X^{''}$. We concatenate $X^{'}$ and $X^{''}$ in the batch size dimension and forward the concatenated batch. Additionally, channel-wise dropout is applied, reducing the dependence on specific channels. These two operations introduce variations and reduce the risk of overfitting in the source domain.

Hierarchical Consistency. Hierarchical consistency promotes the learning of domain-invariant features while encouraging the model to be invariant to domain-specific information. In addition, since perturbation operations are disabled in the inference process, hierarchical consistency also mitigates the gap between training and inference [23]. Specifically, three predictions P_i ($i \in \{1, 2, 3\}$) are produced by multiscale features. As the forwarded input is a concatenation of $X^{'}$ and $X^{''}$, P_i is divided into two parts $P_i^{'}$ and $P_i^{''}$, which corresponds to the same input X. Consistency loss $\mathcal{L}_{con}^i$ is applied to enforce consistency in predictions of the same input. We utilize hierarchical consistency $\mathcal{L}_{con}$ for deep supervision. $\mathcal{L}_{con}^i$ and $\mathcal{L}_{con}$ are calculated as follows:

$$\mathcal{L}_{con}^i = \frac{1}{2}(\mathcal{KL}(P_i^{'}||P_i^{''}) + \mathcal{KL}(P_i^{''}||P_i^{'})), \tag{1}$$

$$\mathcal{L}_{con} = \frac{1}{3}\sum_{i=1}^{3}\mathcal{L}_{con}^i, \tag{2}$$

where $\mathcal{KL}(\cdot)$ denotes Kullback-Leibler divergence [25]. The overall training loss $\mathcal{L}$ is calculated as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{dice} + \alpha\mathcal{L}_{con}, \tag{3}$$

where $\mathcal{L}_{ce}$ denotes cross-entropy loss, $\mathcal{L}_{dice}$ denotes Dice loss, and hyper-parameter α denotes the weight of $\mathcal{L}_{con}$.

3 Experiment

3.1 Dataset and Implementation Details

Experiments are conducted on three cross-domain SDG settings: cross-modality Abdominal MRI-CT [14,15] from T2-SPIR MRI to CT, cross-modality Abdominal CT-MRI [14,15] from CT to T2-SPIR MRI, and cross-sequence Cardiac bSSFP-LGE [26] from bSSFP MRI to LGE MRI. Abdominal T2-SPIR MRI contains 20 3D volumes and Abdominal CT contains 30 3D volumes. Cardiac bSSFP and Cardiac LGE each contain 45 3D volumes. For each setting, the model is trained on the single source domain and tested on the unseen target domain. The initial dataset split follows [6,7]. In this paper, to perform experiments in SDG with extremely few annotations, we utilize only 20%, 20%, and 8% (3, 4, and 3 3D volumes) of the initial training set to train the model on Abdominal MRI-CT, Abdominal CT-MRI, and Cardiac bSSFP-LGE respectively.

Implementation details are listed as follows. Experiments of our method are implemented by PyTorch and NVIDIA RTX A6000. We utilize the AdamW optimizer with the initial learning rate of 0.0003 and apply cosine decay. The batch size is set as 20. For all three settings, the model is trained for 1500 epochs. We evaluate our method at the final epoch with Dice score (%). Hyper-parameter α in Eq. (3) is set as 15. For the proposed perturbation consistency training strategy, intensity perturbation (*e.g.,* Brightness) [12] is adopted and the dropout rate of channel-wise dropout is set as 0.5.

3.2 Experimental Results

We compare the experimental results of our proposed MEDU with existing advanced methods (U-Net [16,9], CSDG [6], AugSeg [12], SLAug [7], and SAM2-UNet [13]) to prove the effectiveness on all three settings. Geometry transformations [6] are employed on all the methods as the common preprocessing operations. Table 1, 2, and 3 summarize the results on Abdominal MRI-CT, Abdominal CT-MRI, and Cardiac bSSFP-LGE respectively. MEDU outperforms all the other methods on both class and average results, as evaluated by the Dice score. Specifically, for Abdominal MRI-CT and Abdominal CT-MRI, MEDU obtains the Dice score of 81.75% and 85.51%, gaining improvements of 12.60% and 4.19% respectively. For Cardiac bSSFP-LGE, MEDU obtains the Dice score of 78.24%. Visualization results are shown in Fig. 3. Source and GT denote the source domain images and ground truths respectively.

Table 1: Comparison of results (%) on Abdominal MRI-CT.

Method	Proportion	Volumes	**Abdominal MRI-CT**				
			Liver	R-Kidney	L-Kidney	Spleen	**Average**
U-Net	20%	3	75.17	12.65	6.47	5.92	25.05
CSDG	20%	3	74.53	64.80	71.76	51.78	65.72
AugSeg	20%	3	81.70	63.48	68.39	55.22	67.20
SLAug	20%	3	79.87	68.22	71.81	56.71	69.15
SAM2-UNet	20%	3	82.74	44.89	28.20	55.69	52.88
MEDU	20%	3	**88.86**	**79.28**	**79.01**	**79.87**	**81.75**

Table 2: Comparison of results (%) on Abdominal CT-MRI.

Method	Proportion	Volumes	**Abdominal CT-MRI**				
			Liver	R-Kidney	L-Kidney	Spleen	**Average**
U-Net	20%	4	27.66	17.35	13.52	22.37	20.23
CSDG	20%	4	62.28	80.81	80.26	56.76	70.03
AugSeg	20%	4	78.26	77.39	81.50	73.73	77.72
SLAug	20%	4	84.11	83.41	81.39	76.38	81.32
SAM2-UNet	20%	4	52.66	66.34	42.09	47.03	52.03
MEDU	20%	4	**87.94**	**88.87**	**85.25**	**79.98**	**85.51**

Table 3: Comparison of results (%) on Cardiac bSSFP-LGE.

Method	Proportion	Volumes	**Cardiac bSSFP-LGE**			
			LVC	MYO	RVC	**Average**
U-Net	8%	3	47.35	15.24	22.60	28.39
CSDG	8%	3	47.11	23.32	33.30	34.58
AugSeg	8%	3	81.64	56.34	66.34	68.11
SLAug	8%	3	86.59	69.71	75.05	77.12
SAM2-UNet	8%	3	73.79	19.01	51.98	48.26
MEDU	8%	3	**87.11**	**71.32**	**76.28**	**78.24**

3.3 Ablation Study

As shown in Table 4, ablation study is conducted on all three settings to demonstrate the effectiveness of two main components (the proposed dual-encoder U-shaped network and hierarchical consistency) in MEDU. $\text{MEDU}^{*}_{\text{CNN}}$ denotes the dual-encoder U-shaped network that employs only CNN-based encoder and incorporates intensity perturbation. $\text{MEDU}^{*}_{\text{Trans}}$ denotes the dual-encoder U-shaped network that employs only Transformer-based encoder and incorporates intensity perturbation. $\text{MEDU}^{*}_{\text{Dual}}$ denotes our MEDU without hierarchical consistency. Specifically, $\text{MEDU}^{*}_{\text{Dual}}$ obtains the Dice score of 81.41%, 85.34%, and 77.86% on three settings respectively, exceeding performances of

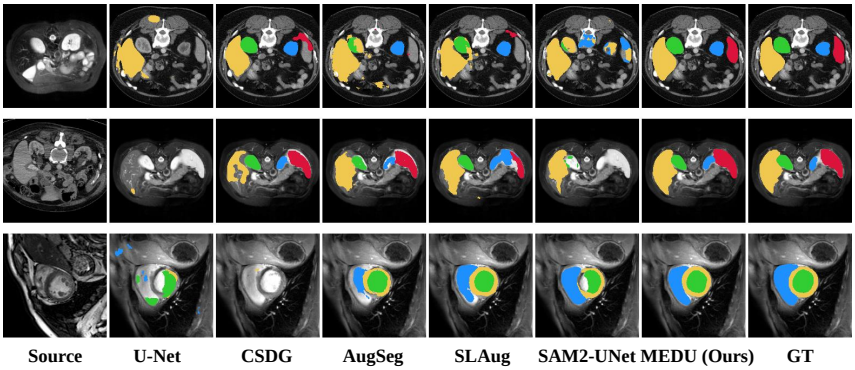

Fig. 3: Visualization results for Abdominal MRI-CT (top row), Abdominal CT-MRI (middle row), and Cardiac bSSFP-LGE (bottom row) respectively.

Table 4: Ablation study on MEDU. Reported results are the average Dice score (%) of classes on three settings. HC denotes hierarchical consistency.

Method	HC	Abdominal MRI-CT	Abdominal CT-MRI	Cardiac bSSFP-LGE
$\text{MEDU}^{*}_{\text{CNN}}$	✗	76.17	83.59	77.82
$\text{MEDU}^{*}_{\text{Trans}}$	✗	81.01	82.97	77.70
$\text{MEDU}^{*}_{\text{Dual}}$	✗	81.41	85.34	77.86
MEDU	✓	**81.75**	**85.51**	**78.24**

both $\text{MEDU}^{*}_{\text{CNN}}$ and $\text{MEDU}^{*}_{\text{Trans}}$. Results here prove that the dual-encoder U-shaped network takes advantage of U-Net and SAM2 encoders while fusing extracted features to learn representative semantic features. Besides, MEDU outperforms $\text{MEDU}^{*}_{\text{Dual}}$ on all three settings, which proves the effectiveness of hierarchical consistency. Hierarchical consistency encourages predictions of the same inputs to remain consistent under various perturbations, aiding in the learning of domain-invariant features and enhanced generalization capability.

4 Conclusion

Distribution shifts and label scarcity are two critical problems that hinder effective medical image segmentation in clinical applications. Thus, we concentrate on single-source domain generalization with extremely few annotations in this paper, which trains the model with extremely few labeled samples in one source domain and aims for generalizable performances in the unseen target domain. A medical dual-encoder framework (MEDU) is proposed to learn domain-invariant features under limited supervision, incorporating a dual-encoder U-shaped network and a perturbation consistency training strategy. A dual-encoder U-shaped network introduces two different encoders and feature fusion modules to learn

representative features, utilizing SAM2 for a good model initialization and helping ease overfitting in the source domain. A perturbation consistency training strategy leverages perturbation operations and hierarchical consistency to enhance generalization capability. Extensive experiments on three cross-domain settings prove the effectiveness of our MEDU.

Acknowledgments. This work is supported by NSFC Project (62222604, 62206052), China Postdoctoral Science Foundation (2024M750424), Fundamental Research Funds for the Central Universities (020214380120, 020214380128), State Key Laboratory Fund (ZZKT2024A14, ZZKT2025B05), Postdoctoral Fellowship Program of CPSF (GZC20240252), Jiangsu Funding Program for Excellent Postdoctoral Talent (2024ZB 242) and Jiangsu Science and Technology Major Project (BG2024031).

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

References

1. Alexander H Berger, Laurin Lux, Nico Stucki, Vincent Bürgin, Suprosanna Shit, Anna Banaszak, Daniel Rueckert, Ulrich Bauer, and Johannes C Paetzold. Topologically faithful multi-class segmentation in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–731. Springer, 2024.
2. Junjie Liang, Peng Cao, Wenju Yang, Jinzhu Yang, and Osmar R Zaiane. 3d-sautomed: Automatic segment anything model for 3d medical image segmentation from local-global perspective. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer, 2024.
3. Qiang Qiao, Wenyu Wang, Meixia Qu, Kun Su, Bin Jiang, and Qiang Guo. Medical image segmentation via single-source domain generalization with random amplitude spectrum synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 435–445. Springer, 2024.
4. Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.
5. Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.
6. Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022.
7. Zixian Su, Kai Yao, Xi Yang, Kaizhu Huang, Qiufeng Wang, and Jie Sun. Rethinking data augmentation for single-source domain generalization in medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2366–2374, 2023.
8. Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2022.

9. Ziqi Zhou, Lei Qi, and Yinghuan Shi. Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In *European Conference on Computer Vision*, pages 420–436. Springer, 2022.
10. Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021.
11. Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random convolutions for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10312–10322, 2023.
12. Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11350–11359, 2023.
13. Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*, 2024.
14. B Landman, Z Xu, J Igelsias, M Styner, T Langerak, and A Klein. Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge (2015). *DOI: https://doi. org/10.7303/syn3193805*, 2015.
15. A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.
16. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.
17. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
18. 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*, pages 574–584, 2022.
19. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
20. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.
21. Xian Lin, Yangyang Xiang, Li Zhang, Xin Yang, Zengqiang Yan, and Li Yu. Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. *arXiv preprint arXiv:2309.06824*, 2023.
22. Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024.

23. Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905, 2021.
24. Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.
25. Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
26. Xiahai Zhuang, Jiahang Xu, Xinzhe Luo, Chen Chen, Cheng Ouyang, Daniel Rueckert, Victor M Campello, Karim Lekadir, Sulaiman Vesal, Nishant RaviKumar, et al. Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge. *Medical Image Analysis*, 81:102528, 2022.