HARP: Harmonization and Adaptive Refinement of Pseudo-Labels for Cross-Domain Medical Image Segmentation

Yulong Liu[1,2], Wenqing Ye[1,2], Hui Liu[2], Ziyi Chen[1,2], Peilin Li[2], Ronald X. Xu[1,2], and Mingzhai Sun[1,2]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230026, P.R.China
xux@ustc.edu.cn, mingzhai@ustc.edu.cn
[2] Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, 215123, P.R.China

Abstract. Medical image segmentation is crucial for accurate diagnosis and effective treatment planning. However, in cross-domain semi-supervised segmentation, the scarcity of labeled data often leads to suboptimal performance and poor generalization across diverse medical imaging domains. Moreover, pseudo-labels generated from unlabeled data are inherently noisy, introducing confirmation bias that destabilizes training and hinders the model's ability to accurately capture complex anatomical structures. To address these challenges, we propose HARP: Harmonization and Adaptive Refinement of Pseudo-Labels for Cross-Domain Medical Image Segmentation, a framework designed to enhance segmentation performance by integrating two novel modules: the Adaptive Pseudo-label Selection (APS) module and the Cross-Domain Harmonization (CDH) module. The APS module ensures the quality and reliability of pseudo-labels by using a confidence-based filtering mechanism and an iterative refinement strategy. The CDH module uses matrix decomposition to harmonize differences across medical imaging modalities, enhancing data diversity while preserving domain-specific features and improving the model's adaptability to varying imaging protocols for robust performance across diverse medical datasets. Extensive experiments on three medical datasets demonstrate the effectiveness of HARP, achieving significant improvements across multiple evaluation metrics. The source code is available at https://github.com/lbllyl/HARP.

Keywords: Semi-supervised learning · Medical image segmentation · Cross-domain adaptation

1 Introduction

Accurate segmentation of anatomical structures is crucial for clinical applications, enabling precise diagnosis and treatment planning [23]. However, acquir-

R. Xu and M. Sun are co-corresponding authors.

ing large-scale annotated medical imaging datasets is challenging due to the need for expert labeling, which is time-consuming and costly [19]. Cross-domain semi-supervised segmentation (CD-SSS) has emerged as a promising solution, leveraging a small amount of labeled data alongside a larger pool of unlabeled data across multiple imaging domains [10,13]. Yet, it faces two major challenges: (1) domain shifts [27] caused by variations in imaging modalities, acquisition protocols, and patient demographics, and (2) noisy supervision signals from unlabeled data [28], which introduce confirmation bias and destabilize training.

Recent efforts in semi-supervised and cross-domain medical image segmentation have made progress in addressing these issues. For example, EPL [26] improves pseudo-label quality through Fourier transformations, while MiDSS [16] mitigates domain shifts by constructing intermediate domains. ABD [8] enhances consistency learning via bidirectional patch displacement, and AstMatch [30] uses adversarial self-training to improve cross-domain consistency. Generic [21] focuses on distribution-invariant features using an Aggregating & Decoupling framework. Despite these advancements, existing methods often struggle with the complexity and variability of medical data, leading to models that may overfit or fail to generalize in real-world clinical settings.

To tackle these challenges, we propose HARP: Pseudo-Labeling with Adaptive Selection and Harmonization for Cross-Domain Medical Image Segmentation. HARP introduces two key components: an Adaptive Pseudo-label Selection (APS) module and a Cross-Domain Harmonization (CDH) module. The APS module improves pseudo-label reliability by filtering low-confidence predictions and refining labels iteratively, ensuring stable and accurate supervision. The CDH module addresses domain gaps by aligning feature distributions across domains using Singular Value Decomposition (SVD), while preserving modality-specific characteristics. By synthesizing hybrid images that blend features from different domains, the CDH module enhances the model's robustness to domain variability.

HARP effectively combines pseudo-label refinement and domain harmonization to address the dual challenges of limited labeled data and domain shifts. The APS module ensures high-quality pseudo-labels, reducing noise and confirmation bias [9,15], while the CDH module enriches the training set with hybrid images, improving generalization [25,20]. Extensive experiments on three medical datasets demonstrate HARP's effectiveness, achieving significant improvements [12,6,1] in segmentation accuracy, showcasing HARP's potential as a reliable and generalizable solution for cross-domain medical image segmentation in clinical practice.

2 Method

Problem Setting. Cross-Domain Semi-Supervised Segmentation (CD-SSS) seeks to perform accurate segmentation across N medical imaging domains $\{D_i\}_{i=1}^{N}$ with limited labeled data $\{(x_i, y_i)\}_{i=1}^{K}$ and a larger unlabeled set $\{u_i\}_{i=1}^{M}$ ($M \gg K$). Each image has resolution $H \times W \times D$ and multi-class labels $y_i \in \{0,1\}^{H \times W \times C}$,

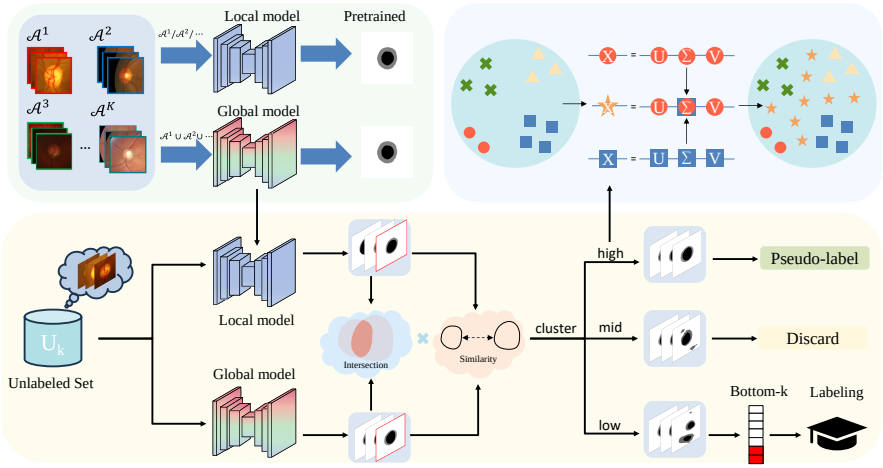

Fig. 1. Overview of the proposed HARP framework. The HARP framework processes labeled and unlabeled data from multiple domains to improve segmentation performance. The local model is trained sequentially on data from one domain at a time, while the global model is periodically updated and used to reinitialize the local model. The APS module refines pseudo-labels generated from unlabeled data, while the CDH module performs data mixup to create augmented samples.

where C is the number of classes. The task is to leverage limited labeled data $\{(x_i, y_i)\}_{i=1}^{K}$ and abundant unlabeled data $\{u_i\}_{i=1}^{M}$ ($M \gg K$) from N domains $\{D_i\}_{i=1}^{N}$ to achieve accurate segmentation across diverse medical imaging domains.

2.1 Overall Pipeline

The HARP framework trains segmentation models across N domains iteratively. Initially, a local model is trained on each domain using domain-specific data. These local models are then aggregated into a global model, which is trained on data from all domains. The global model's parameters are shared back to the local models for further refinement, and this process repeats until global training is complete. Finally, the global model is fine-tuned on domain-specific data to produce N specialized local models.

The Adaptive Pseudo-label Selection (APS) module improves pseudo-labels through confidence filtering and iterative refinement, while the Cross-Domain Harmonization (CDH) module reduces domain gaps by aligning features using Singular Value Decomposition (SVD). As illustrated in Fig. 1, the global model is further trained using labeled data, refined pseudo-labels, and harmonized data, resulting in a robust final model.

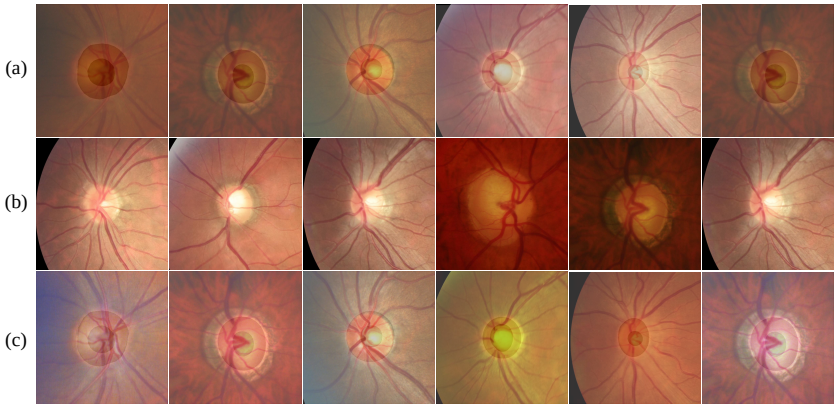

Fig. 2. Visual results from the Fundus dataset. (a) Labeled training data. (b) Unlabeled data. (c) Data processed by the CDH module, showing increased diversity and inherited masks.

2.2 Adaptive Pseudo-label Selection Module

The Adaptive Pseudo-label Selection Module selects high-quality pseudo-labels by combining predictions from a global model f_u and N fine-tuned local models f_d. For unlabeled data in each domain, f_d and f_u independently generate pseudo-labels $\hat{y}_d$ and $\hat{y}_u$, respectively. These pseudo-labels are then evaluated for reliability using a confidence score $C(\hat{y}_d, \hat{y}_u)$, which integrates two key metrics: the Intersection score and a Fréchet-based similarity measure.

The confidence score is defined as:

$$C(\hat{y}_d, \hat{y}_u) = \underbrace{\frac{2\sum_{i=1}^{N} \hat{y}_{d,i} \cdot \hat{y}_{u,i}}{\sum_{i=1}^{N} \hat{y}_{d,i} + \sum_{i=1}^{N} \hat{y}_{u,i}}}_{\text{Intersection score } D(\hat{y}_d, \hat{y}_u)} \times \underbrace{\left(1 - \frac{F(\hat{y}_d, \hat{y}_u)}{L}\right)}_{\text{Similarity measure } S(\hat{y}_d, \hat{y}_u)}, \tag{1}$$

where $\hat{y}_{d,i}$ and $\hat{y}_{u,i}$ are binary values indicating whether the i-th pixel belongs to a specific class, $F(\hat{y}_d, \hat{y}_u)$ is the Fréchet distance, which measures the similarity between the two sets of pseudo-labels. It is defined as:

$$F(\hat{y}_d, \hat{y}_u) = \inf_{\alpha,\beta} \max_{t \in [0,1]} d\left(\hat{y}_d(\alpha(t)), \hat{y}_u(\beta(t))\right), \tag{2}$$

where α and β are continuous, non-decreasing reparameterizations of the pseudo-labels, and d is the distance metric, L is the image diagonal length, used to normalize the Fréchet distance.

The Intersection score $D(\hat{y}_d, \hat{y}_u)$ quantifies the overlap between the pseudo-labels, with higher values indicating greater agreement between the local and global models. The similarity measure $S(\hat{y}_d, \hat{y}_u)$ evaluates their spatial alignment, ranging from 0 to 1, where higher values denote greater similarity. The confidence

score $C(\hat{y}_d, \hat{y}_u)$ (Eq. 1) combines these metrics, providing a robust indicator of pseudo-label quality.

To categorize the confidence scores, we employ the k-means clustering algorithm, grouping the scores into high, medium, and low confidence:

$$\arg \min_{\{C_1,C_2,C_3\}} \sum_{i=1}^{3} \sum_{x \in C_i} \|x - \mu_i\|^2, \quad \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x. \tag{3}$$

Here, C_1, C_2, and C_3 represent the clusters for high, medium, and low confidence, respectively. T_{large} represents the average of C_1 minimum confidence and C_2 maximum confidence, while T_{small} is the average of C_2 minimum confidence and C_3 maximum confidence. Pseudo-labels with confidence scores above the threshold $T = \max(T_{\text{large}}, 1 - T_{\text{small}})$ are retained as reliable data.

To address class imbalance, the annotation budget for each domain is set to $\frac{B_k}{N}$, where k denotes different domains with B as the total annotation budget. The module selects the $\frac{B_k}{N}$ pseudo-labels with the lowest confidence scores for manual annotation, incorporating them into the training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}, y_i \in \mathcal{Y}$, where y_i is the ground-truth label for x_i. By prioritizing low-confidence samples, the module ensures that the most challenging and informative examples are included in the training process, optimizing the use of limited annotation resources.

2.3 Cross-Domain Harmonization Module

The Cross-Domain Harmonization Module aims to reduce domain gaps and enhance data diversity by aligning features across different domains. Given two input images, $X_1 \in \mathbb{R}^{m \times n}$ and $X_2 \in \mathbb{R}^{m \times n}$, the images are defined as follows:

$$X_1 \in D_{\text{train}}^{\text{labeled}} \cup D_{\text{train}}^{\text{annotated}}, \quad X_2 \in D_{\text{unlabeled}}^{\text{first } \beta}. \tag{4}$$

Here, β represents the proportion of unlabeled data from the target domain that is selected for the harmonization process. These images originate from different domains, and the module begins by performing SVD independently:

$$X_1 = U_1 \Sigma_1 V_1^T \quad , \quad X_2 = U_2 \Sigma_2 V_2^T, \tag{5}$$

where $U_1, U_2 \in \mathbb{R}^{m \times m}$ and $V_1, V_2 \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\Sigma_1, \Sigma_2 \in \mathbb{R}^{m \times n}$ are diagonal matrices containing the singular values of X_1 and X_2, respectively.

The module then proceeds to mix the singular values of the two images using a mixup operation. To control the interpolation strength, a parameter α is drawn from a Beta distribution, which controls the interpolation strength. The mixed singular value matrix Σ_{mix} is computed as follows:

$$\Sigma_{\text{mix}} = \alpha \Sigma_1 + (1 - \alpha) \Sigma_2. \tag{6}$$

To reconstruct the augmented image X_{mix}, the module employs the mixed singular value matrix Σ_{mix} along with the left and right singular vectors of the source domain image:

$$X_{\text{mix}} = U_1 \Sigma_{\text{mix}} V_1^T. \tag{7}$$

The reconstructed image X_{mix} inherits the structural information from the source domain image while incorporating the semantic content from both domains through the mixed singular values. This process effectively bridges the domain gap and enhances the model's ability to generalize across diverse datasets, as shown in Fig. 2. Regarding the ground truth labels for the augmented images, the module directly copies the labels from the source domain image, which ensures that the label information remains consistent with the original source domain data, preserving the integrity of the training process.

By operating in the latent space of singular values (Eq. 5), the module captures the essential characteristics of the images from both domains, enabling effective domain adaptation. The mixup operation on the singular values (Eq. 6) allows for a smooth interpolation between domains, generating augmented samples that bridge the gap between the source and target distributions.

2.4 Optimization

The Cross-Domain Harmonization Module is optimized using a combined loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}, \tag{8}$$

where $\mathcal{L}_{\text{CE}}$ measures classification error, and $\mathcal{L}_{\text{Dice}}$ evaluates the overlap between predicted and ground truth segmentation masks. This combined loss ensures accurate classification and high-quality segmentation, enabling the model to generalize effectively across domains.

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluated our method on three medical imaging datasets: Fundus [22], Prostate [14], and M&MS [4]. Fundus contains 789 retinal images from 4 centers, Prostate includes 1,510 MRI slices from 6 sources, and M&MS comprises 3,447 cardiac CT slices from 4 sources. For training, we used all unlabeled data, 20 labeled samples (40 samples on Prostate), and an additional 5% labeled budget. Performance was evaluated using Dice Score, Jaccard Score (both in %), 95% Hausdorff Distance (95HD, in voxels), and Average Surface Distance (ASD, in voxels).

Table 1. Comparisons with SSL and AL methods on fundus, prostate, and M&MS datasets.

Dataset	Type	Method	DC ↑	JC ↑	HD ↓	ASD ↓
Fundus	SSL	BCP	83.05	73.66	11.05	5.80
		CauSSL	61.81	51.80	41.25	23.94
	AL	RLD	71.89	63.59	24.51	16.53
		FEAL	73.03	63.50	18.97	11.56
	SSAL	**Ours**	**88.22**	**80.26**	**7.81**	**3.74**
Prostate	SSL	BCP	64.81	55.17	52.60	27.22
		CauSSL	20.93	15.48	114.62	73.30
	AL	RLD	59.15	50.86	34.98	15.87
		FEAL	79.19	70.15	20.83	8.82
	SSAL	**Ours**	**81.38**	**72.81**	**18.13**	**8.69**
M&MS	SSL	BCP	71.65	62.67	30.91	18.22
		CauSSL	35.44	26.73	72.90	37.99
	AL	RLD	77.90	69.19	18.68	8.79
		FEAL	80.85	72.53	11.30	5.48
	SSAL	**Ours**	**85.10**	**76.93**	**6.74**	**3.03**

3.2 Implementation Details

The proposed method is implemented using the PyTorch framework and trained on an NVIDIA GeForce RTX 3090 GPU. The hyperparameter β is set to 10, and the optimization is performed using the Adam optimizer with an initial learning rate of 1×10^{-4}. Other parameters remain consistent with [16].

3.3 Main Results

Our experiments (Table 1) across three medical imaging datasets—Fundus, Prostate, and M&MS—demonstrate the versatility and effectiveness of our method. We compare against four state-of-the-art methods: two Semi-Supervised Learning (SSL) approaches (BCP [2] and CauSSL [17]) and two Active Learning (AL) approaches (RLD [18] and FEAL [7]), as shown in Table 1.

On the Fundus dataset, our method achieved a Dice Coefficient (DC) of 88.22 and Jaccard Coefficient (JC) of 80.26, outperforming both SSL and AL methods. The nearest competitor, BCP, trailed by over 5 percentage points. Our method also excelled in boundary precision, with a Hausdorff Distance (HD) of 7.81 and Average Surface Distance (ASD) of 3.74. This strong performance is particularly notable given the challenging dual-object segmentation task and the limited labeled data (only 20 samples).

For the Prostate dataset, our method again delivered superior results, with a DC of 81.38 and JC of 72.81. It handled complex prostate boundaries effectively,

achieving an HD of 18.13 and ASD of 8.69. While SSL methods struggled with the variability in prostate shapes and intensities, even the better-performing AL methods fell short of our results.

On the M&MS dataset, which requires precise segmentation of the left ventricle, myocardium, and right ventricle, our method achieved a DC of 85.10 and JC of 76.93. Boundary precision was particularly strong, with an HD of 6.74 and ASD of 3.03. Compared to existing approaches, our method consistently achieves higher accuracy and robustness, making it well-suited for clinical applications where precision is critical.

Table 2. Ablation study of different components on two datasets

Components			**Fundus Dataset**				**M&MS Dataset**			
L	APS	CDH	DC ↑	JC ↑	HD ↓	ASD ↓	DC ↑	JC ↑	HD ↓	ASD ↓
✓			54.53	42.44	52.11	28.16	43.08	33.82	56.53	32.47
✓	✓		76.48	67.09	21.66	12.80	70.82	60.67	24.67	11.04
✓		✓	84.73	76.18	14.05	6.52	82.24	73.72	8.88	5.11
✓	✓	✓	**88.22**	**80.26**	**7.81**	**3.74**	**85.10**	**76.93**	**6.74**	**3.03**

3.4 Ablation Studies

Table 2 summarizes our ablation study, demonstrating the contribution of each HARP component on the Fundus and M&MS datasets. We evaluated performance using four metrics: Dice Coefficient (DC), Jaccard Index (JC), Hausdorff Distance (HD), and Average Surface Distance (ASD) [23,19].

In the table, L indicates the use of labeled data during training. On the Fundus dataset, the baseline model achieves a DC of 54.53 and JC of 42.44. Adding the Adaptive Pseudo-label Selection Module (APS) significantly improves performance, with DC rising to 76.48 and JC to 67.09. The full model, which combines L, APS, and the Cross-Domain Harmonization Module (CDH), achieves the best results: DC of 88.22 and JC of 80.26, along with notable reductions in HD and ASD [5,11].

Similar improvements are observed on the M&MS dataset. The baseline model yields a DC of 43.08 and JC of 33.82. With APS, these values increase to 70.82 and 60.67, respectively. The full model further boosts performance, reaching a DC of 85.10 and JC of 76.93, while also improving HD and ASD [9,15].

Qualitatively, our method outperforms RLD [18] and FEAL [7] in segmenting Fundus images, producing results that align more closely with ground truth. Similar advantages are seen on the Prostate dataset, where our approach captures prostate regions more accurately [29]. On the M&MS dataset, our method consistently captures fine anatomical details, demonstrating its robustness across diverse medical imaging tasks [24,3].

4 Conclusion

In this study, we introduced the Adaptive Domain Integration Framework (HARP), a novel approach for cross-domain semi-supervised medical image segmentation. HARP efficiently utilizes limited labeled data by training both domain-specific and universal models [9,15]. The Adaptive Pseudo-label Selection Module (APS) enhances the reliability of pseudo-labels, improving training under minimal supervision, while the Cross-Domain Harmonization Module (CDH) mitigates domain shift using Singular Value Decomposition to increase data variability and reduce disparities [5,11]. Both APS and CDH are plug-and-play components, allowing easy integration into existing frameworks [24,3]. Our experiments on three public medical datasets demonstrate that HARP outperforms existing methods across four key metrics [23,19], representing a significant advancement in medical image segmentation.

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

References

1. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds (2020), https://arxiv.org/abs/1906.03671
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation (2023), https://arxiv.org/abs/2305.00673
3. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19786–19797 (2023)
4. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
5. Blanchard, G., Deshmukh, A.A., Dogan, U., Lee, G., Scott, C.: Domain generalization by marginal transfer learning. Journal of machine learning research **22**(2), 1–55 (2021)
6. Cao, Y.T., Shi, Y., Yu, B., Wang, J., Tao, D.: Knowledge-aware federated active learning with non-iid data (2023), https://arxiv.org/abs/2211.13579
7. Chen, J., Ma, B., Cui, H., Xia, Y.: Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts (2024), https://arxiv.org/abs/2312.02567
8. Chi, H., Pang, J., Zhang, B., Liu, W.: Adaptive bidirectional displacement for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4070–4080 (2024)
9. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. IEEE transactions on medical imaging **39**(8), 2626–2637 (2020)

10. Fan, J., Liu, D., Chang, H., Huang, H., Chen, M., Cai, W.: Taxonomy adaptive cross-domain adaptation in medical imaging via optimization trajectory distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21174–21184 (2023)
11. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
12. Kim, S., Bae, S., Song, H., Yun, S.Y.: Re-thinking federated active learning based on inter-class diversity (2023), https://arxiv.org/abs/2303.12317
13. Li, J., Zhu, G., Hua, C., Feng, M., BasheerBennamoun, Li, P., Lu, X., Song, J., Shen, P., Xu, X., Mei, L., Zhang, L., Shah, S.A.A., Bennamoun, M.: A systematic collection of medical image datasets for deep learning (2021), https://arxiv.org/abs/2106.12864
14. Litjens, G., Toth, R., Van De Ven, W., Hoeks, C., Kerkstra, S., Van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al.: Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis **18**(2), 359–373 (2014)
15. Lyu, F., Ye, M., Carlsen, J.F., Erleben, K., Darkner, S., Yuen, P.C.: Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. IEEE Transactions on Medical Imaging **42**(3), 797–809 (2022)
16. Ma, Q., Zhang, J., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11642–11651 (2024)
17. Miao, J., Chen, C., Liu, F., Wei, H., Heng, P.A.: Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21426–21437 (October 2023)
18. Song, J., Kim, T.S., Kim, J., Nam, G., Kooi, T., Choo, J.: Is user feedback always informative? retrieval latent defending for semi-supervised domain adaptation without source data (2024), https://arxiv.org/abs/2407.15383
19. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation (2020), https://arxiv.org/abs/1908.10454
20. Triantafillou, E., Larochelle, H., Zemel, R., Dumoulin, V.: Learning a universal template for few-shot dataset generalization. In: International Conference on Machine Learning. pp. 10424–10433. PMLR (2021)
21. Wang, H., Li, X.: Towards generic semi-supervised framework for volumetric medical image segmentation. Advances in Neural Information Processing Systems **36** (2024)
22. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Transactions on Medical Imaging **39**(12), 4237–4248 (2020)
23. Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L.: Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Medical Image Analysis **55**, 88–102 (Jul 2019). https://doi.org/10.1016/j.media.2019.04.005, http://dx.doi.org/10.1016/j.media.2019.04.005
24. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15651–15660 (2023)

25. Xu, Z., Li, W., Niu, L., Xu, D.: Exploiting low-rank structure from latent domains for domain generalization. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13. pp. 628–643. Springer (2014)
26. Yao, H., Hu, X., Li, X.: Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 3099–3107 (2022)
27. Zhang, L., Wu, J., Wang, L., Wang, L., Steffens, D.C., Qiu, S., Potter, G.G., Liu, M.: Brain anatomy-guided mri analysis for assessing clinical progression of cognitive impairment with structural mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer (2023)
28. Zhao, Z., Wang, Z., Wang, L., Yu, D., Yuan, Y., Zhou, L.: Alternate diverse teaching for semi-supervised medical image segmentation. In: European Conference on Computer Vision. pp. 227–243. Springer (2024)
29. Zhou, Y., Huang, J., Wang, C., Song, L., Yang, G.: Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21085–21096 (2023)
30. Zhu, G., Zhang, J., Liu, J., Du, X., Hao, R., Liu, Y., Liu, L.: Astmatch: Adversarial self-training consistency framework for semi-supervised medical image segmentation. arXiv preprint arXiv:2406.19649 (2024)