Retrieval-Guided Semi-Supervised Federated Learning for Medical Image Segmentation

Heejung Park[1], Soopil Kim[2], Wafa Arsalane[1], Sion An[3†], and Sang Hyun Park[3,4†]

[1] Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Republic of Korea
[2] AX Research Group for Semiconductors, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Republic of Korea
[3] Graduate School of Artificial Intelligence, Pohang Unversity of Science and Technology (POSTECH), Pohang, Republic of Korea
[4] Department of Computer Science & Engineering, Pohang Unversity of Science and Technology (POSTECH), Pohang, Republic of Korea
sionan624@postech.ac.kr, sanghyunpark@postech.ac.kr
† Co-corresponding authors.

Abstract. Semi-supervised federated learning (SSFL) has emerged as a promising approach for medical image segmentation, enabling collaborative learning across institutions using both limited labeled and abundant unlabeled data without sharing sensitive patient information. However, existing methods utilize publicly available labeled data merely as a generic training pool for knowledge transfer, failing to exploit the fine-grained semantic information embedded in individual public samples that could directly inform the segmentation of unlabeled local data. To overcome this limitation, we propose a **R**etrieved **R**eference-guided **S**emi-**S**upervised **Fed**erated segmentation framework **(R2S2Fed)** that explicitly leverages public labeled data as contextual references rather than a generic knowledge source. During local training, our method dynamically retrieves semantically relevant public references for each unlabeled local sample based on feature similarity, and leverages their representations to generate query-specific segmentation parameters through a dynamic segmentation head. This provides explicit supervision for unlabeled local data, enabling each client to perform adaptive segmentation tailored to its local data characteristics. Extensive experiments demonstrate that our proposed method consistently outperforms existing SSFL approaches in both IID and Non-IID settings. The code will be made publicly available upon publication.

Keywords: Semi-supervised federated learning · Medical image segmentation · Retrieval.

1 Introduction

Although federated learning (FL) [10,11,12,21,22,27] addresses privacy concerns by enabling decentralized training without sharing data, it does not resolve the annotation burden yet. Expert labeling in medical imaging requires specialized knowledge that is both costly and time-consuming, leaving the majority of clinical data unlabeled across institutions [9,19]. Thus, semi-supervised federated learning (SSFL) [5] has emerged as a promising solution to tackle this challenge by exploiting both labeled and unlabeled data during collaborative training. SSFL paradigms are generally categorized based on the distribution of labeled and unlabeled data into two settings: (1) *labels-for-private*, where each client holds both labeled and unlabeled data [7,14,15,16,24], and (2) *labels-for-public*, where labels reside exclusively in a publicly available dataset at the server, while clients maintain only private unlabeled data [13,20,26]. In this study, we specifically focus on the *labels-for-public* setting, which presents a highly practical scenario with the increasing availability of public medical datasets.

Under this setting, recent studies have focused on transferring knowledge from the public dataset to decentralized clients through model-driven approaches such as pseudo-labeling and parameter alignment. For example, SemiFL [5] performs supervised learning on the global server while local clients learn from unlabeled data via pseudo-labels. Building on this, $(FL)^2$ [13] and BSemiFL [26] improve pseudo-label reliability through client-specific confidence thresholding and adaptive global-local model integration, respectively. Ma *et al.* [20] address model homogeneity and heterogeneity in SSFL via LSSL and HSSF by incorporating a self-assessment module for prediction confidence estimation. However, these methods inevitably compress public data into generalized model parameters or predictions, losing the fine-grained semantic information of individual samples that could otherwise directly inform local segmentation.

In this paper, we propose a novel **R**etrieved **R**eference-guided **S**emi-**S**upervised **Fed**erated segmentation framework **(R2S2Fed)**, explicitly leveraging the semantic representations of references retrieved from public data. Our key insight is that public data contains semantically similar data for local client data, which can provide fine-grained guidance for accurate predictions. Unlike previous methods that utilize public data solely as a source of global knowledge, R2S2Fed retrieves relevant references for each local data and employ a dynamic weight generator that produces query-specific weights, enabling focused learning on the most relevant semantics and handling data heterogeneity across clients.

The main contributions of this work are as follows:

1. We propose a retrieved reference-guided semi-supervised federated segmentation framework that utilizes publicly available labeled data as query-conditioned references rather than a source of global knowledge transfer, enabling sample-level adaptive inference for each client.
2. Within this framework, we introduce a query-adaptive retrieval that retrieves relevant references to condition a dynamic parameter generator, producing query-specific segmentation parameters for each client data.
3. Through extensive experiments on independent and identically distributed (IID) and non-IID medical image benchmarks, we demonstrate that R2S2Fed significantly outperforms the state-of-the-art semi-supervised federated learning methods.

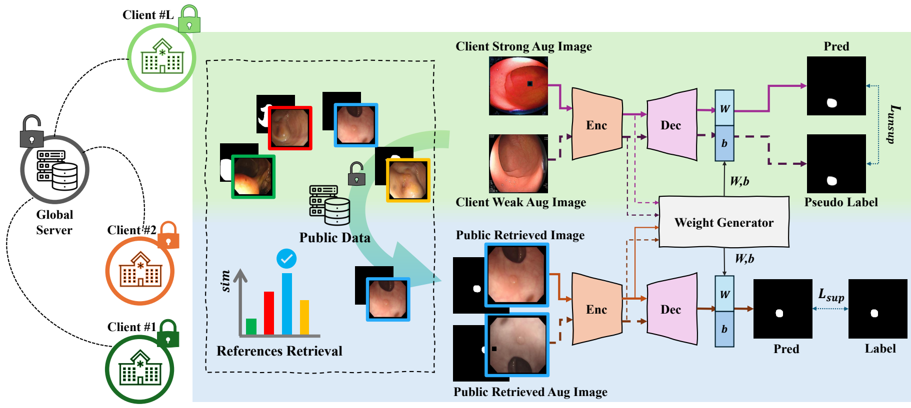

Fig. 1: Illustration of the local training structure of R2S2Fed framework. For each client's unlabeled image, references are retrieved from a public labeled dataset via feature similarity and used to guide prediction through a dynamic weight generator. Unsupervised consistency loss is applied across two augmented views of the unlabeled image, while supervised loss is imposed on augmented views of the retrieved reference images. Aggregation is performed at a global server following the federated learning protocol.

2 Method

2.1 Problem settings

The goal of SSFL is to collaboratively optimize a global model leveraging both labeled public data and unlabeled local client data without directly sharing local client data. Thus, we assume that the existence of a publicly available dataset $\mathcal{D}^P$ is accessible to both the global server and all local clients, and a local client dataset $\mathcal{D}^L$ is privately held by each local client L. $\mathcal{D}^P$ comprises N_P pairs of an image x_i^P and a segmentation mask y_i^P. $\mathcal{D}^L$ consists of unlabeled images without corresponding segmentation masks.

2.2 Reference-Guided Dynamic Segmentation Framework

Our framework is inspired by the hypernetwork-based framework [18,29] that utilizes semantic information of the target to facilitate task-agnostic segmentation. Building on this, we propose a reference-guided dynamic segmentation framework that leverages semantic representations from references to adaptively segment the query image, as illustrated in Fig. 1. Given an unlabeled client image $x^Q \in \mathcal{D}^L$, the model takes a set of K relevant references $\mathcal{R} = \{(x_i^R, y_i^R)\}_{i=1}^K \subset \mathcal{D}^P$. Both x^Q and $\mathcal{R}$ are projected into a shared latent space through an encoder ϕ, yielding feature representations $f^Q = \phi(x^Q)$ and $f_i^R = \phi(x_i^R)$. The foreground and background features $F^R = [F_f\,;\,F_b] \in \mathbb{R}^{2\times C}$ are obtained by performing masked average pooling on each f_i^R with its annotation y_i^R and subsequently averaging the results across all K references, yielding F_f and $F_b \in \mathbb{R}^{1\times C}$. To align the reference semantics with the query feature, F^R and $f^Q \in \mathbb{R}^{BHW\times C}$ are processed through a cross-attention module *i.e.*, F^R is used as the query and f^Q as the key and value in the attention module. This enables the model to selectively extract query-relevant semantics from the references. The resulting representation $\omega \in \mathbb{R}^{2\times C}$ directly defines the parameters of the dynamically generated segmentation head, where the foreground channel serves as convolutional weight and the background channel is averaged into a scalar bias. The query feature f^Q is then decoded and passed through the segmentation head to produce the final mask for x^Q. This allows each client to perform query-adaptive segmentation, effectively leveraging semantics in references beyond knowledge transferring.

Query-Adaptive Retrieval. The effectiveness of the proposed framework depends on selecting references that provide guidance for each query. To this end, we introduce a query-adaptive retrieval (QAR) to dynamically construct $\mathcal{R}$ by selecting semantically relevant representations from $\mathcal{D}^P$.

Static or randomly sampled references [28,29] fail to consider the semantic evolution of query features or domain shifts that may occur throughout training. Consequently, the progressively updated encoder can become misaligned with reference representations selected without considering the current feature space, leading to degraded semantic correspondence over time.

To mitigate this issue, QAR dynamically retrieves query-relevant references based on the evolving query features to promote continuous semantic alignment during training. Specifically, at every T communication rounds, we recompute feature embeddings for the query and all public data $\mathcal{D}^P = \{(x_j^P, y_j^P)\}_{i=1}^{N_P}$ using the encoder at t^{th} communication rounds parameterized by θ_t *i.e.*, $f_t^Q = \phi_{\theta_t}(x^Q)$ and $f_{j,t}^P = \phi_{\theta_t}(x_j^P)$. To find the references, we compute the L2 distance between f_t^Q and $f_{j,t}^P$ and perform Top-K retrieval. The corresponding images and masks from these indices constitute the references $\mathcal{R}$, providing query-relevant semantic context. This strategy ensures that the retrieved references remain semantically aligned with the query, promoting robustness to client heterogeneity and consistent segmentation guidance.

Training Strategy During local client training, we employ FixMatch [25], which combines supervised learning on the references with consistency regularization on unlabeled client data.

For the retrieved $\mathcal{R}$, we apply strong augmentation on the data and predict the foreground regions using the original images as corresponding references. The supervised loss on references is defined as: $\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{CE}}(\hat{y}^R, y^R) + \mathcal{L}_{\text{IoU}}(\hat{y}^R, y^R)$.

For consistency regularization, pseudo-labels $\tilde{y}^Q$ are generated from weakly augmented predictions and filtered by a confidence threshold τ. The unsupervised loss $\mathcal{L}_{\text{unsup}}$ is computed as a combination of cross-entropy and Dice loss over pixels exceeding τ. The total loss is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda\mathcal{L}_{\text{unsup}}$, where λ balances the supervised and unsupervised terms.

After local training, client models are aggregated at the server using FedAvg [22]. Following aggregation, the server performs additional training on the public dataset $\mathcal{D}^P$ to further stabilize the global model as commonly practiced in SSFL [5,26]. To enhance robustness and ensure diverse reference exposure, reference images are randomly selected rather than retrieved based on similarity. The updated model is then transferred to clients for the next round.

3 Experimental Settings

Datasets and Evaluation Metrics We evaluate R2S2Fed on two medical image segmentation tasks: polyp segmentation and skin lesion segmentation. To assess performance under different data heterogeneity conditions, we conduct experiments in both IID and non-IID settings. We use Dice coefficient (Dice) and Intersection over Union (IoU) to report the results. For the IID setting, we use ISIC2018 [3] dataset for skin lesion segmentation. We randomly partition the 2,694 training images into five groups consisting of a public dataset with 100 images and four clients with 200, 400, 800, and the remaining images, followed by [20]. For the non-IID setting, we use four polyp segmentation datasets. We assign CVC-ColonDB [2] with 380 images as the public dataset and CVC-ClinicDB [1] with 612 images, NeoPolyp [23] with 1,000 images, and Kvasir [8] with 1,000 images to three clients. In both settings, each client's data is partitioned into 80% for training and 20% for testing. All images are resized to 380×380.

Implementation Details. We follow the encoder-decoder architecture following [20] and ResNet34 [6] is employed as the encoder backbone. For data augmentation, we adopt the weak augmentation strategy from [20,25] consisting of random horizontal flip and small random crops, and a strong augmentation including Randaugment [4] with color jittering, rotation, affine transformations. To focus the model on reference semantics, masks are randomly reversed during training. Pretraining is conducted for initial 20 epochs on the public dataset $\mathcal{D}^P$ at the global server, as in [20], followed by 180 rounds of federated training where each client performs one local epoch per communication round. The threshold τ starts at 0.5 and is increased by 0.05 per communication round, up to a maximum of 0.9. All models are optimized using AdamW optimizer [17] with a learning rate of 1×10^{-4} and weight decay of 0.02. We use $K = 4$ references for each query and retrieves for every $T = 10$ communication rounds. The balance weight λ for the unsupervised loss is set to 1.0.

Comparison Methods. We compare R2S2Fed with five semi-supervised federated learning methods: SemiFL [5], LSSL [20], $(\text{FL})^2$ [13] and BSemiFL [26] with homogeneous models and HSSF [20] with heterogeneous models. For fair comparison, all methods are trained with the same data splits, augmentation

Table 1: Segmentation results (Dice/IoU) on ISIC and poly segmentation datasets.

Dataset		Client	SemiFL	LSSL	HSSF	$(\mathrm{FL})^2$	BSemiFL	R2S2Fed
IID	ISIC	# 1	**86.52/79.38**	76.43/64.11	85.48/77.69	85.98/78.20	84.70/77.61	86.34/78.90
		# 2	83.33/74.44	75.72/64.59	83.16/74.57	82.85/74.41	83.32/75.24	**84.14/75.56**
		# 3	83.19/75.71	76.97/65.54	82.29/73.87	85.17/77.03	80.94/73.53	**85.47/77.53**
		# 4	81.61/72.78	76.10/63.75	80.28/70.77	84.83/75.87	82.64/74.02	**85.07/76.29**
		Avg.	83.66/75.58	76.30/64.50	82.80/74.34	84.71/76.38	82.90/75.10	**85.26/77.07**
Non-IID	Polyp	# 1	60.81/52.58	62.50/50.14	58.99/48.64	66.32/57.48	61.13/51.89	**75.71/66.80**
		# 2	54.97/46.27	54.03/43.17	53.94/45.08	47.63/39.67	52.45/44.51	**62.54/53.54**
		# 3	50.89/41.17	62.44/50.19	53.59/43.59	49.22/39.29	40.65/32.67	**63.21/51.96**
		Avg.	55.56/46.68	59.66/47.83	55.51/45.77	54.39/45.48	51.41/43.02	**67.15/57.43**

strategies, and communication rounds. Homogeneous models use ResNet34 as the encoder. For HSSF with heterogeneous models, we employ two ResNet18 and one or two ResNet34 encoders for clients, and ResNet101 for the global server.

4 Experimental Results

SSFL on homogeneous datasets. To evaluate R2S2Fed in homogeneous setting, we compare the performance on the ISIC2018 [3], where all clients share the same data distribution. As shown in upper rows of Table 1, R2S2Fed outperforms all other methods with the highest average performance of 85.26% Dice and 77.07% IoU score, achieving the best performance on three out of four clients and remaining competitive on the client 1. In contrast, SemiFL, HSSF and BSemiFL exhibit performance drops on certain clients. This confirms that dynamically retrieving and leveraging relevant references enhance the robustness of segmentation performance across clients.

SSFL on heterogeneous datasets. We further investigate the effectiveness of R2S2Fed in heterogeneous setting by comparing performance on the polyp segmentation datasets, which built from four institutions. As shown in the lower rows of Table 1, we observe a similar tendency with IID settings *i.e.*, achieving the best performance across all clients. Existing SSFL methods struggle under data heterogeneity, showing substantial performance degradation compared to IID results. In particular BSemiFL achieves only 51.41% Dice score, the lowest among all comparison methods, indicating that Bayesian weighting alone is insufficient to handle severe domain shift. SemiFL, which shows competitive performance on homogeneous setting, fails to capture heterogeneous data distributions across clients due to its reliance on a global model for pseudo-label generation. HSSF underscores LSSL by 4.15% despite using a larger ResNet101 encoder, suggesting that distilling knowledge from domain-mismatched public data becomes ineffective regardless of model capacity. LSSL achieves the second

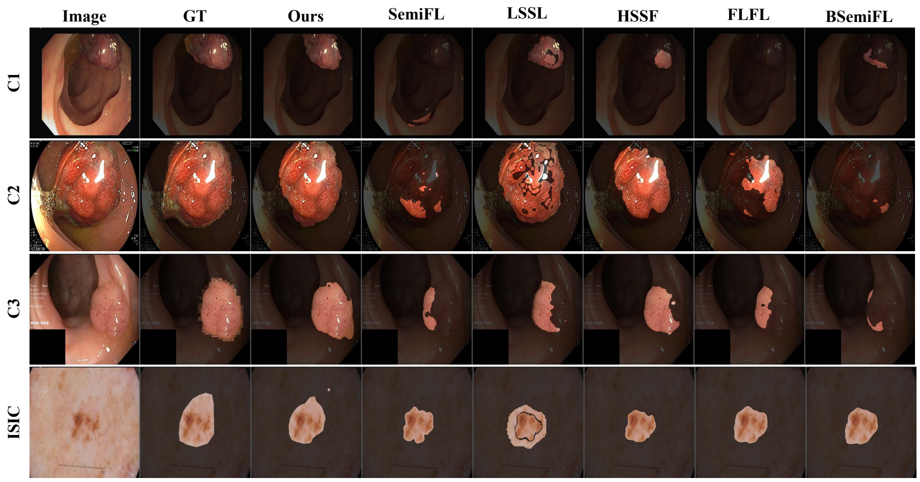

Fig. 2: Qualitative comparison of segmentation results under heterogeneous (rows 1–3) and homogeneous (row 4) settings. Columns from left to right: input image, ground truth (GT), and predictions of the proposed and comparison methods.

best average performance, yet still falls short in front of R2S2Fed by 7.49% in Dice and 9.60% in IoU. This demonstrates that our retrieved reference-guided segmentation framework effectively handles severe data heterogeneity by selecting relevant references tailored to each client's data.

Qualitative comparison. Figure 2 shows comparative qualitative results from both IID and non-IID settings. Each row displays an input image, ground truth mask and predictions from different methods. The first three rows each client's results in non-IID setting, while the last row presents results from the IID setting on the ISIC dataset. While comparison methods fail to accurately identify lesion or produce complete segmentation masks, R2S2Fed consistently achieves better segmentation on both datasets. It demonstrates that our reference-guided segmentation framework successfully utilizes semantically similar references, leading to more robust predictions across diverse data distributions.

Ablation Study for QAR. We first examine the impact of the number of retrieved references by varying K from 1 to 8. As shown in Table 2a, K between 2 and 4 achieves superior performance across both datasets, though the optimal value varies depending on dataset characteristics. For the ISIC dataset, where client and public images share the same domain distribution, retrieval quality remains stable up to $K = 4$, allowing more reference images to contribute positively to performance. In contrast, the substantial domain gap in the polyp dataset limits semantically relevant retrievals. As shown in Figure 3, while ISIC retrievals exhibit consistent appearance across all four references, only one or

Table 2: Ablation study of the QAR. All results are presented as Dice score (%)

(a) Effect of the retrieval size

K	1	2	4	8
ISIC	83.67	85.18	85.26	84.60
Poly	68.68	68.09	67.15	62.45

(b) Impact of retrieval on train and test

Train	×	×	✓	✓
Test	×	✓	×	✓
Avg	62.42	62.43	67.13	67.15

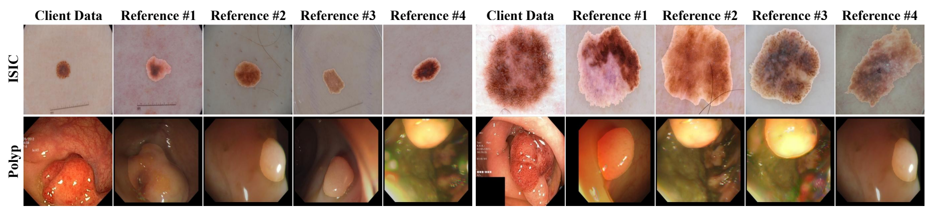

Fig. 3: Visualization of top-4 retrieved public images with overlaid labels for representative client queries from ISIC (top) and Polyp (bottom) datasets.

two retrieved images show meaningful similarity to the client data in the polyp dataset. Nevertheless, R2S2Fed on polyp achieves its best performance at $K = 1$, demonstrating that our method can extract meaningful guidance from a single highly relevant image under large domain shifts. Notably, both datasets exhibit performance degradation at $K = 8$, suggesting that beyond a certain threshold, less relevant samples are included in the reference set, diluting its overall quality.

We further investigate the role of retrieval by comparing four settings where retrieval is applied or omitted at each of the training and inference phases. As shown in Table 2b, when random references are used during training, performance degrades by approximately 5% regardless of whether retrieval at inference. In contrast, the performance gain from applying retrieval at inference is relatively marginal compared to the gain from retrieval at training when relevant references are used. These results demonstrate that semantically relevant references during training play a critical role in guiding the dynamic parameter generation, enabling the model to consistently learn query-relevant semantic representations throughout the training process.

5 Conclusion

In this paper, we proposed a retrieved reference-guided framework for semi-supervised federated learning in a setting where only public data are labeled while client data remain unlabeled. The proposed Query-Adaptive Retrieval (QAR) dynamically selects semantically relevant public references for each client's query and leverages them to generate query-specific segmentation weights, enabling instance-level adaptation. This design facilitates semantic alignment between public and client data, enhancing learning stability on unlabeled data.

Extensive experiments in both IID and non-IID settings demonstrate consistent performance improvements over existing semi-supervised federated learning methods. Future work will focus on exploring more scalable retrieval strategies and extending this framework to other medical image analysis tasks.

Acknowledgments. This work is supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea government (MOTIE) (No. RS-2025-02633871), the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439264, Development of High-Performance Machine Unlearning Technologies for Privacy Protection), National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-00516124), and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (No.RS-2025-25433877).

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

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
2. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition **45**(9), 3166–3182 (2012)
3. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
4. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
5. Diao, E., Ding, J., Tarokh, V.: Semifl: Semi-supervised federated learning for unlabeled clients with alternate training. Advances in Neural Information Processing Systems **35**, 17871–17884 (2022)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Jeong, W., Yoon, J., Yang, E., Hwang, S.J.: Federated semi-supervised learning with inter-client consistency and disjoint learning. arXiv preprint arXiv:2006.12097 (2020)
8. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International conference on multimedia modeling. pp. 451–462. Springer (2019)
9. Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., Jin, C.: Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Computers in Biology and Medicine **169**, 107840 (2024)

10. Kang, M., Chikontwe, P., Kim, S., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H.: Efficient one-shot federated learning on medical data using knowledge distillation with image synthesis and client model adaptation. Medical Image Analysis p. 103714 (2025)
11. Kim, S., Park, H., Chikontwe, P., Kang, M., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H.: Communication efficient federated learning for multi-organ segmentation via knowledge distillation with image synthesis. IEEE Transactions on Medical Imaging (2025)
12. Kim, S., Park, H., Kang, M., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H.: Federated learning with knowledge distillation for multi-organ segmentation with partially labeled datasets. Medical image analysis **95**, 103156 (2024)
13. Lee, S., Le, T.L.V., Shin, J., Lee, S.J.: FL^2: Overcoming few labels in federated semi-supervised learning. Advances in Neural Information Processing Systems **37**, 43693–43714 (2024)
14. Liang, X., Lin, Y., Fu, H., Zhu, L., Li, X.: Rscfed: Random sampling consensus federated semi-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10154–10163 (2022)
15. Liu, Q., Yang, H., Dou, Q., Heng, P.A.: Federated semi-supervised medical image classification via inter-client relation matching. In: International conference on medical image computing and computer-assisted intervention. pp. 325–335. Springer (2021)
16. Liu, Y., Shang, X., Zhang, Y., Lu, Y., Gong, C., Xue, J.H., Wang, H.: Mind the gap: Confidence discrepancy can guide federated semi-supervised learning across pseudo-mismatch. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10173–10182 (2025)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Lu, Z., He, S., Zhu, X., Zhang, L., Song, Y.Z., Xiang, T.: Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8741–8750 (2021)
19. Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 318–329. Springer (2021)
20. Ma, Y., Wang, J., Yang, J., Wang, L.: Model-heterogeneous semi-supervised federated learning for medical image segmentation. IEEE Transactions on Medical Imaging **43**(5), 1804–1815 (2024)
21. Mächler, L., Ezhov, I., Kofler, F., Shit, S., Paetzold, J.C., Loehr, T., Zimmer, C., Wiestler, B., Menze, B.H.: Fedcostwavg: A new averaging for better federated learning. In: International MICCAI Brainlesion Workshop. pp. 383–391. Springer (2021)
22. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
23. Ngoc Lan, P., An, N.S., Hang, D.V., Long, D.V., Trung, T.Q., Thuy, N.T., Sang, D.V.: Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In: International Symposium on Visual Computing. pp. 15–28. Springer (2021)

24. Saha, P., Mishra, D., Noble, J.A.: Rethinking semi-supervised federated learning: How to co-train fully-labeled and fully-unlabeled client imaging data. In: International conference on medical image computing and computer-assisted intervention. pp. 414–424. Springer (2023)
25. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
26. Wang, H., Wang, S., Li, J., Ren, H., Han, X., Xu, W., Guo, S., Zhang, T., Li, R.: Bsemifl: Semi-supervised federated learning via a bayesian approach. In: Proceedings of the Forty-second International Conference on Machine Learning (2025)
27. Wu, N., Yu, L., Yang, X., Cheng, K.T., Yan, Z.: Fediic: Towards robust federated learning for class-imbalanced medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 692–702. Springer (2023)
28. Zhao, L., Chen, X., Chen, E.Z., Liu, Y., Chen, T., Sun, S.: Retrieval-augmented few-shot medical image segmentation with foundation models. IEEE Transactions on Neural Networks and Learning Systems (2025)
29. Zhao, X., Pang, Y., Ji, W., Sheng, B., Zuo, J., Zhang, L., Lu, H.: Spider: A unified framework for context-dependent concept segmentation. In: International Conference on Machine Learning. pp. 60906–60926. PMLR (2024)