Learnable In-Context Templates for Medical Image Segmentation via In-Context Learning

Xueqi Bao, Chang Zheng, Chengwen Zhang, and Qicheng Lao*

Beijing University of Posts and Telecommunications, Beijing, China
xueqi.bao@bupt.edu.cn, qicheng.lao@bupt.edu.cn

Abstract. In-Context Learning has emerged as a promising paradigm for automated medical image segmentation, enabling foundation models like SegGPT to generalize to new tasks using annotated support examples. However, existing approaches typically rely on retrieving raw images as prompts, which suffer from two critical limitations: the inherent noise in raw medical data often misleads the model, and the repetitive retrieval process incurs severe computational latency. To address these challenges, we propose a novel framework, Learnable In-Context Templates, which shifts the paradigm from sample retrieval to template learning. We introduce an In-Context Distillation strategy to optimize a compact set of learnable templates, condensing the representative lesion patterns of a target dataset. By distilling essential semantic features and filtering out raw image noise, these refined templates provide a cleaner, more concentrated support set. Extensive experiments on six datasets demonstrate that our method consistently outperforms state-of-the-art ICL baselines. Notably, our approach shows broad adaptability across diverse medical modalities and achieves a substantial order-of-magnitude reduction in inference latency, offering an efficient solution for clinical deployment. Code will be available at: https://github.com/MembrAI/In-Context-Templates.

Keywords: Medical Image Segmentation · In-Context Learning · Template Generation

1 Introduction

Medical image segmentation is a cornerstone of clinical diagnosis and treatment planning [14,28]. Recent advances in vision foundation models have introduced In-Context Learning (ICL) as a promising paradigm for automated segmentation [3]. Instead of updating model parameters for new tasks, ICL conditions segmentation on a set of annotated support examples, where each example consists of both the medical image and its corresponding segmentation mask. These image–mask pairs serve as structured visual prompts, explicitly conveying the target anatomy or lesion pattern to be segmented. Representative frameworks such as SegGPT [22] and UniverSeg [4] demonstrate that exemplar-guided prompting enables flexible and training-free adaptation across diverse segmentation tasks.

Building upon this philosophy, current state-of-the-art ICL methods [9,23,20] typically adopt a *retrieve-and-prompt* strategy: for each query image, semantically similar support examples are retrieved from a training pool based on feature similarity. While effective, we argue that this instance-level retrieval paradigm suffers from two fundamental limitations. First, raw medical images are inherently noisy and heterogeneous; retrieving the nearest neighbor sample does not guarantee a high-quality contextual prompt [26,8], as irrelevant background details or ambiguous lesion boundaries in the retrieved samples can introduce confounding signals, thereby degrading segmentation performance. Second, retrieval incurs substantial computational overhead [5,16], as searching large databases and loading high-resolution support images significantly increases inference latency, limiting practicality in real-time or resource-constrained clinical settings.

To overcome these limitations, we revisit a key question: Is explicit instance retrieval truly necessary for in-context medical image segmentation? Although medical objects exhibit considerable visual variability, they often share intrinsic morphological regularities that can be abstracted into representative semantic priors. This observation motivates a paradigm shift from *instance-level matching* to *template-level abstraction*. Inspired by prompt-based fine-tuning [13,15,2] and prototype learning [25,27,7], we propose **Learnable In-Context Templates**, a compact set of trainable representations that encode core lesion morphologies at a higher level of abstraction. In contrast to retrieval-based approaches [9,23,20], which rely on selecting raw exemplar instances as contextual support, our framework directly learns structured semantic templates, thereby avoiding the redundancy, noise, and computational overhead inherent to sample retrieval.

Furthermore, to address the potential rigidity of fixed templates, we propose an *In-Context Template Distillation* strategy to adaptively refine the templates. Rather than serving as static priors, the templates are progressively optimized to capture a comprehensive spectrum of principal lesion patterns while suppressing irrelevant background artifacts and spurious correlations. This distillation process enables the templates to serve as a cleaner, more representative, and more transferable contextual support compared to retrieved instances.

We conduct extensive experiments on six diverse medical segmentation benchmarks, and results demonstrate that our approach consistently outperforms state-of-the-art ICL baselines. Moreover, by eliminating the retrieval bottleneck, our framework reduces inference latency by two orders of magnitude, achieving a substantially improved trade-off between segmentation accuracy and computational efficiency—an essential requirement for real-world clinical deployment.

2 Methods

2.1 Overview

To address the bottlenecks of noise susceptibility and prohibitive retrieval latency in in-context medical image segmentation, we propose an In-Context Template Learning framework. As illustrated in Fig. 1, the framework comprises three

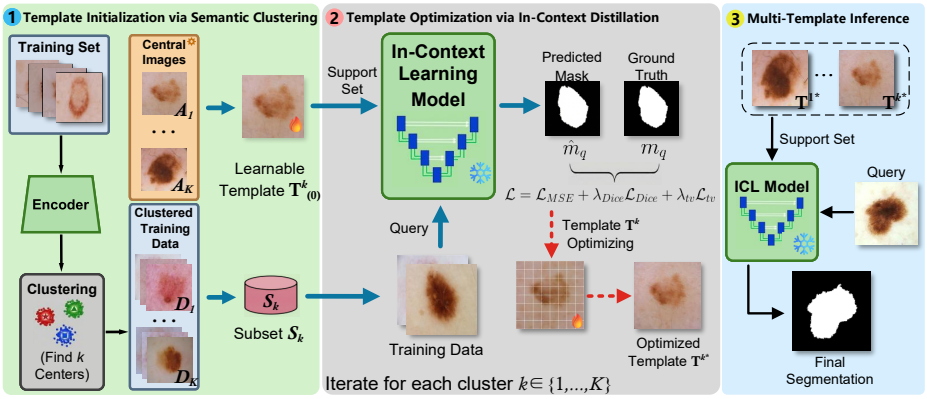

Fig. 1. Overview of the proposed In-Context Templates Learning framework.

primary stages: (1) Template Initialization: This stage establishes initial representative templates by partition the training set into semantic clusters based on extracted visual features; (2) Template Optimization: We iteratively optimize the learnable templates via in-context distillation, effectively filtering out raw image noise to capture the representative semantic features of the target lesions; (3) Multi-Template Inference: During the testing phase, the framework utilizes all learned templates to construct a comprehensive support set, providing robust, noise-free contextual guidance for unseen queries.

Unlike prior efforts that select, tune, or condense existing data representations, our method learns explicit image–mask templates as reusable in-context support. By preserving lesion-level structures and reducing reliance on noisy raw exemplars, these templates are well suited for medical ICL segmentation.

2.2 Template Initialization via Semantic Clustering

To initialize a set of appropriate templates capable of capturing the diverse lesion morphologies within the dataset (e.g., polyps or skin lesions with varying sizes and textures), we first perform semantic grouping over the training set, clustering samples into multiple semantically coherent subspaces.

Specifically, given training set $\mathcal{D}_{train} = \{x_i, m_i\}_{i=1}^{N}$ where x_i denotes the original medical image, and m_i represents the corresponding segmentation mask, we utilize a pre-trained image encoder Enc of vision foundation models, e.g., SAM [12], to extract high-dimensional feature embeddings $\{f_i\}_{i=1}^{N}$ for all training images. We then employ the K-Means algorithm to partition the feature space into K semantic clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$. To accelerate model convergence and provide a representative initial prior, for the k-th cluster, we calculate the geometric centroid of each cluster and formulate its corresponding initial template, formally defined as $\mathbf{T}^k_{(0)} = (\hat{x}^k_{(0)}, \hat{m}^k_{(0)})$, where $\hat{x}^k_{(0)}$ and $\hat{m}^k_{(0)}$ represent the center image of $\mathcal{C}_k$ and its corresponding mask. This initialization is achieved by

selecting the geometric centroid-matched image-mask pair:

$$\mathbf{T}_{(0)}^{k} = \arg \min_{(x,m)\in\mathcal{C}_k} ||Enc(x) - \mu_k||_2, \tag{1}$$

where μ_k denotes the mean feature of the k-th cluster.

Crucially, this initialization step concurrently establishes a localized distillation environment for each template. Having anchored the initial template $\mathbf{T}_{(0)}^{i}$, we define the remaining samples within the corresponding subspace as a specific training subset, $\mathcal{S}_k = \mathcal{C}_k \setminus \{\mathbf{T}_{(0)}^{k}\}$. Each subset $\mathcal{S}_k$ is exclusively dedicated to training and fine-tuning its corresponding k-th Learnable Template. Through this strategy, we decompose a complex global optimization problem into K relatively simpler local fitting tasks. This ensures that each template focuses solely on learning a specific lesion morphology pattern within the data.

2.3 Template Optimization via In-Context Distillation

This section elucidates the core of our proposed framework: In-Context Template Distillation. Traditional fine-tuning paradigms typically necessitate updating massive model parameters θ, which is prone to triggering overfitting and catastrophic forgetting in medical scenarios characterized by scarce annotated samples [11,24]. To address this challenge, drawing inspiration from Prompt Tuning in large language models [13,10,19], we propose a parameter-efficient training strategy: we maintain the parameters θ of the segmentation model frozen, while treating the input prompt as a learnable continuous variable.

Building upon the initialized template set $\{\mathbf{T}_{(0)}^{k}\}_{k=1}^{K}$ derived in Sec. 2.2, our objective is to optimize each $\mathbf{T}^k$, transforming it from a static, raw instance into a template-based representation that encapsulates the representative semantic features of its corresponding subset.

During the training phase, optimization is performed independently for each template. For the k-th cluster, we sequentially sample data from the training subset $\mathcal{S}_k$ as query inputs, i.e., $(x_q, m_q) \sim \mathcal{S}_k$. We utilize the learnable template $\mathbf{T}^k = (\hat{x}^k, \hat{m}^k)$ as the context prompt and feed it into the frozen pre-trained in-context learning model $\mathcal{F}_\theta$. The model performs inference on the query image guided by the visual cues provided by the template:

$$\hat{m}_q = \mathcal{F}_\theta(x_q, \mathbf{T}^k) = \mathcal{F}_\theta(x_q, \mathbb{S} = \{(\hat{x}^k, \hat{m}^k)\}), \tag{2}$$

where $\mathbb{S}$ denotes the support set. During this process, $\mathbf{T}^k$ participates in the computation as a learnable tensor, enabling it to capture subset-level characteristics and generalize into a template representation. Compared with the raw cluster-center exemplars used for initialization, the optimized templates are task-adapted prompts that preserve useful lesion morphology while suppressing sample-specific noise and background artifacts.

Optimization Objective To guide the evolution of the templates, we design a hybrid loss function, incorporating both segmentation fidelity and visual constraint terms. We concurrently employ Mean Squared Error and Dice loss [17]

to measure the discrepancy between the predicted mask $\hat{m}_q$ and the ground truth m_q. MSE focuses on pixel-level intensity consistency; Dice loss targets global geometric overlap:

$$\mathcal{L}_{seg} = \mathcal{L}_{MSE}(\hat{m}_q, m_q) + \lambda_{dice}\mathcal{L}_{Dice}(\hat{m}_q, m_q). \tag{3}$$

Since we directly optimize $\hat{x}^k$ in the pixel space, unconstrained optimization tends to degenerate the template into adversarial patterns containing high-frequency noise. Consequently, we incorporate Total Variation loss [18] $\mathcal{L}_{tv}$ to constrain spatial smoothness:

$$\mathcal{L}_{tv}(\hat{x}^k) = \sum_{i,j} \left(|\hat{x}^k_{i+1,j} - \hat{x}^k_{i,j}| + |\hat{x}^k_{i,j+1} - \hat{x}^k_{i,j}| \right). \tag{4}$$

The total loss function is a weighted sum of the segmentation loss and the total variation loss:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{tv}\mathcal{L}_{tv}(\hat{x}^k). \tag{5}$$

Ultimately, we seek the optimal template $\mathbf{T}^{k*}$ for the k-th cluster by minimizing the following objective function via the backpropagation algorithm:

$$\mathbf{T}^{k*} = \arg\min_{\hat{x}^k, \hat{m}^k} \mathbb{E}_{(x_q, m_q)\sim\mathcal{S}_k} \left[\mathcal{L}_{seg}(\mathcal{F}_\theta(x_q, \mathbf{T}^k), m_q) + \lambda_{tv}\mathcal{L}_{tv}(\hat{x}^k) \right]. \tag{6}$$

Following this distillation process, the resulting $\mathbf{T}^{k*}$ not only preserves the core characteristics of the specific lesion category but also filters out extraneous background noise, thereby providing precise contextual guidance during inference.

2.4 Multi-Template Inference

Upon the convergence of the distillation phase, we obtain a set of optimized expert templates $\mathbf{T}^1, \ldots, \mathbf{T}^K$ that collectively encapsulate the diverse semantic patterns of the data distribution. During the inference phase, given an unseen test image x_{test}, we bypass the retrieval process utilized by traditional ICL methods. Conversely, we aggregate all K learned templates to constitute a comprehensive global expert ensemble as the support set, which is jointly fed into the model with the test image:

$$\hat{m}_{test} = \mathcal{F}_\theta(x_{test}, \mathbb{S} = \{\mathbf{T}^{1*}, \ldots, \mathbf{T}^{K*}\}). \tag{7}$$

By incorporating all expert templates, we ensure that the support set covers the full semantic manifold of the target lesions, from simple geometries to complex textures. Leveraging the inherent capacity of in-context learning models to process multi-example contexts, the network dynamically attends to the template features most compatible with x_{test}, thereby achieving robust segmentation on unseen test cases. This formulation targets broad multi-modality adaptability in zero-shot ICL settings rather than explicit cross-dataset domain adaptation.

Table 1. Quantitative comparison of Dice scores (%) across six datasets. The improvements over the Dual-SC baseline are marked in red.

Method	ISIC	TNUI	CVC	BUSI	WBC D1	WBC D2
SAM (auto) [12]	38.61	30.05	19.86	48.02	48.83	53.86
PerSAM [29]	52.42	12.73	16.07	36.44	62.89	56.23
SegGPT [22]	58.47	31.76	66.03	30.96	90.29	90.89
SegGPT+Dual-SC [9]	74.42	68.93	82.51	58.21	91.09	90.70
SegGPT+Ours	78.61 (+4.19)	80.34 (+11.41)	**83.34** (+0.83)	61.90 (+3.69)	**94.57** (+3.48)	**95.08** (+4.38)
UniverSeg [4]	74.41	61.35	44.18	65.30	91.75	89.20
UniverSeg+Dual-SC [9]	81.51	76.03	65.55	71.88	92.71	89.34
UniverSeg+Ours	**83.21** (+1.70)	**83.92** (+7.89)	66.13 (+0.58)	**73.62** (+1.74)	94.17 (+1.46)	91.46 (+2.12)

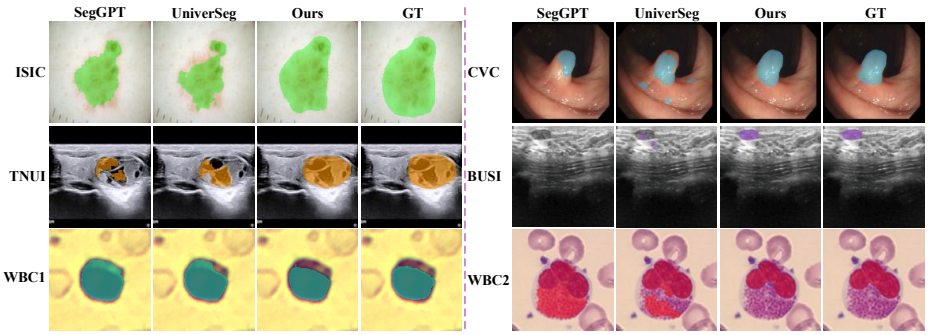

Fig. 2. Qualitative comparison of segmentation results across six datasets.

3 Experiments

3.1 Dataset and Experimental setup

Datasets. We validate the effectiveness of our approach using six diverse medical segmentation datasets: ISIC2018 [6], TNUI2021 [31], CVC-ColonDB [21], BUSI [1], WBC Dataset1 [30], and WBC Dataset2 [30]. All utilized datasets provide high-quality binary segmentation masks for supervision and evaluation.
Method Comparison. To comprehensively evaluate the effectiveness of our proposed method, we benchmark it against several state-of-the-art segmentation approaches. Specifically, we employ SAM (Auto) [12] and PerSAM [29] as foundation models; notably, SAM (Auto) [12] operates in a zero-shot mode without task-specific prompts, serving as a critical baseline to highlight the inherent challenges of unguided segmentation. For generalist ICL models, we compare against SegGPT [22] and UniverSeg [4]. Moreover, we include the Dual Similarity Checkup (Dual-SC) method [9], which represents the leading strategy for

Table 2. Comparison of computational efficiency (measured in ms) for different prompt selection strategies based on SegGPT [22].

Method	ISIC	TNUI	CVC	BUSI
Random	21.45	19.87	20.40	22.13
Dual-SC [9]	48.52	44.52	44.49	42.34
Ours	**2.90**	**2.79**	**2.61**	**2.52**

Table 3. Ablation study on the template optimization (Opt.) and objective functions.

Opt.	Loss			Dice Score (%)			
	$\mathcal{L}_{mse}$	$\mathcal{L}_{dice}$	$\mathcal{L}_{tv}$	ISIC	TNUI	CVC	BUSI
–	–	–	–	62.22	34.39	68.01	25.80
✓	–	✓	✓	69.81	66.49	72.45	50.42
✓	✓	–	✓	73.56	73.75	74.49	55.23
✓	✓	✓	–	77.38	78.41	82.23	59.43
✓	✓	✓	✓	**78.61**	**80.34**	**83.34**	**61.90**

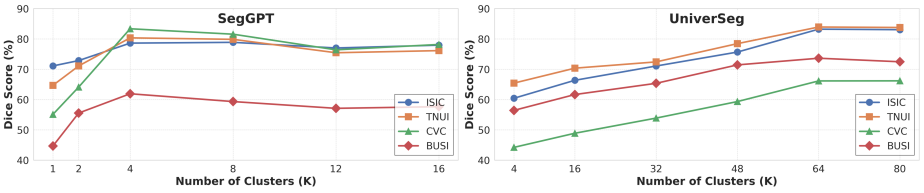

Fig. 3. Ablation study on the effect of cluster number (K).

optimizing in-context example selection.

Implementation Details. Our proposed method is implemented using the PyTorch framework and executed on dual NVIDIA RTX 4090 GPUs. To ensure consistency with the baseline UniverSeg [4], all images are uniformly resized to a resolution of 128×128 and normalized to the $[0, 1]$ intensity range. We evaluate performance using five-fold cross-validation across all datasets. For hyperparameter settings, we set $\lambda_{dice} = 0.2$ and $\lambda_{tv} = 0.001$. We further investigate the selection of λ_{dice} and λ_{tv} in Section 4.4. Regarding comparative experiments with Dual-SC [9], unless otherwise specified, we follow standard protocols where UniverSeg [4] employs a support set size of 64, whereas the SegGPT [22] model utilizes a support set size of 4.

3.2 Experimental Results

Quantitative and Qualitative Results. Table 1 presents the quantitative comparison of our proposed framework against state-of-the-art in-context learning models [22,4] and the advanced retrieval-based Dual-SC method [9] across six datasets. Unlike Dual-SC, which is limited by the specificity and noise of individual raw samples, our optimized templates encapsulate the universal semantic features of the dataset, providing more robust and representative guidance. This advantage is evident on the TNUI dataset, where our method outperforms Dual-SC by 11.41%, confirming that our templates effectively filter out irrelevant details to focus on core lesion patterns. Qualitative results in Fig. 2 align with these metrics, showing that our method generates coherent masks with precise boundaries even in challenging scenarios.

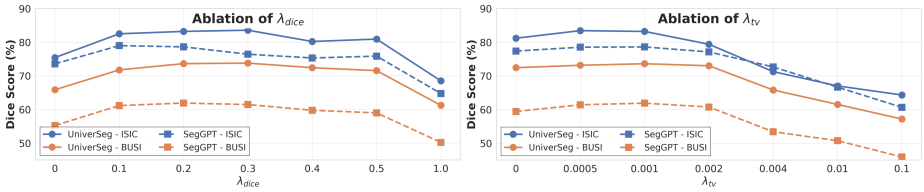

Fig. 4. Ablation study on the effect of λ_{dice} and λ_{tv}.

Efficiency Analysis. We also evaluate the inference efficiency. As shown in Table 2, retrieval-based methods [22,9] suffer from severe latency, primarily bottlenecked by the repetitive I/O overhead of loading and preprocessing support images from the disk. In contrast, our approach is inherently retrieval-free, which eliminates the need for run-time sample search and data loading. Moreover, in practical applications dealing with larger-scale datasets, this efficiency advantage becomes even more pronounced. Template learning introduces a one-time offline optimization cost. However, this cost is incurred only once per dataset, while retrieval-based methods repeatedly perform support-set search, disk loading, and preprocessing for each query. Once optimized, the templates are fixed and reusable across test queries, amortizing the offline cost as evaluation or deployment scales. Thus, inference latency remains independent of the training pool size, with substantially reduced storage I/O.

3.3 Ablation Study

To investigate the contribution of the template optimization strategy and each loss component, we conduct a component-wise ablation study. Table 3 validates the efficacy of our proposed components. Optimizing expert templates via in-context distillation yields a massive performance leap over static baselines, evidenced by the Dice score surging from 34.39% to 80.34% on the TNUI dataset. For loss functions, combining $\mathcal{L}_{mse}$ and $\mathcal{L}_{dice}$ is essential to provide comprehensive supervision for both pixel-level intensity and global geometry. Moreover, the Total Variation constraint ($\mathcal{L}_{tv}$) proves crucial in preventing learnable templates from degenerating into high-frequency noise, ensuring robust generalization.

To ensure a strictly fair comparison with existing works, we align the number of clusters K (i.e., support set size) with standard evaluation protocols: $K = 4$ for SegGPT [22] and $K = 64$ for UniverSeg [4]. As illustrated in Fig. 3, our ablation study on K further validates the rationale behind this alignment. Reducing K below these standard values leads to significant performance degradation, as a limited number of templates cannot adequately cover the diverse morphology of the training distribution. Conversely, increasing K beyond the standard settings yields no further improvement, indicating that performance reaches saturation. Furthermore, as illustrated in Fig. 4, we evaluate the sensitivity of our framework to the loss weights λ_{dice} and λ_{tv} on two datasets. Empirical results demonstrate

that setting $\lambda_{dice} = 0.2$ and $\lambda_{tv} = 0.001$ achieves the optimal balance. Removing these constraints weakens global geometric alignment and structural fidelity, whereas excessively high values severely degrade segmentation accuracy. Thus, this default configuration is uniformly applied across all our experiments.

4 Conclusion

We present a new paradigm for in-context medical image segmentation by replacing inefficient instance-level retrieval with learnable in-context templates. By distilling diverse lesion morphologies into a compact set of expert prototypes as learnable templates, our framework suppresses inherent data noise while eliminating the computational overhead of runtime retrieval. Extensive experiments across six benchmarks demonstrate consistent and state-of-the-art performance. We believe this template-driven strategy paves the way for the efficient and robust deployment of vision foundation models in real-time clinical workflows.

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

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022)
3. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. Advances in neural information processing systems **35**, 25005–25017 (2022)
4. Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universeg: Universal medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21438–21451 (2023)
5. Caccia, L., Ansell, A., Ponti, E., Vulić, I., Sordoni, A.: Training plug-n-play knowledge modules with deep context distillation. arXiv preprint arXiv:2503.08727 (2025)
6. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
7. Feng, Y., Wang, Y., Li, H., Qu, M., Yang, J.: Learning what and where to segment: A new perspective on medical image few-shot segmentation. Medical image analysis **87**, 102834 (2023)
8. Fu, H., Ouyang, Y., Chang, K.W., Wang, Y., Huang, Z., Cai, Y.: Contextnav: Towards agentic multimodal in-context learning. arXiv preprint arXiv:2510.04560 (2025)

9. Gao, J., Lao, Q., Kang, Q., Liu, P., Du, C., Li, K., Zhang, L.: Boosting your context by dual similarity checkup for in-context learning medical image segmentation. IEEE Transactions on Medical Imaging **44**(1), 310–319 (2024)
10. Han, C., Wang, Q., Cui, Y., Wang, W., Huang, L., Qi, S., Liu, D.: Facing the elephant in the room: Visual prompt tuning or full finetuning? arXiv preprint arXiv:2401.12902 (2024)
11. He, A., Wu, Y., Wang, Z., Li, T., Fu, H.: Dvpt: Dynamic visual prompt tuning of large pre-trained models for medical image analysis. Neural Networks **185**, 107168 (2025)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
13. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 3045–3059 (2021)
14. Li, R., Wu, L., Gu, J., Xu, Q., Chen, W., Cai, X., Bu, J.: Moe-sam: Enhancing sam for medical image segmentation with mixture-of-experts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 367–377. Springer (2025)
15. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too. AI open **5**, 208–215 (2024)
16. Mao, R., Li, H., Dong, C., Navab, N., Zhou, S.K.: Randp: Effective and efficient medical visual in-context learning via a retrieve-and-propagate module for prompt-query fusion. In: Medical Imaging with Deep Learning (2026)
17. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
18. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4), 259–268 (1992)
19. Shi, Z., Lipani, A.: Dept: Decomposed prompt tuning for parameter-efficient fine-tuning. arXiv preprint arXiv:2309.05173 (2023)
20. Suo, W., Lai, L., Sun, M., Zhang, H., Wang, P., Zhang, Y.: Visual prompt selection for in-context learning segmentation. In: European Conference on Computer Vision (ECCV) (2024)
21. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017**(1), 4037190 (2017)
22. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 (2023)
23. Wu, C., Restrepo, D., Shuai, Z., Liu, Z., Shen, L.: Efficient in-context medical segmentation with meta-driven visual prompt selection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 255–265. Springer (2024)
24. Xin, Y., Yang, J., Luo, S., Du, Y., Qin, Q., Cen, K., He, Y., Zhang, Z., Fu, B., Yang, X., et al.: Parameter-efficient fine-tuning for pre-trained vision models: A survey and benchmark. arXiv preprint arXiv:2402.02242 (2024)
25. Yin, Z., Xing, E., Shen, Z.: Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. Advances in Neural Information Processing Systems **36**, 73582–73603 (2023)

26. Zhang, J., Wang, B., Liu, H., Nakashima, Y., Nagahara, H.: Panicl: Mitigating over-reliance on single prompt in visual in-context learning. arXiv preprint arXiv:2509.21926 (2025)
27. Zhang, J., Xie, Y., Wang, Y., Xia, Y.: Inter-slice context residual learning for 3d medical image segmentation. IEEE Transactions on Medical Imaging **40**(2), 661–672 (2020)
28. Zhang, M., Dai, X., Sun, Y., Wang, J., Yao, Y., Gong, X., Cong, F., Wang, F., Lv, Y.: Hierarchical self-prompting sam: A prompt-free medical image segmentation framework. arXiv preprint arXiv:2506.02854 (2025)
29. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
30. Zheng, X., Wang, Y., Wang, G., Liu, J.: Fast and robust segmentation of white blood cell images by self-supervised learning. Micron **107**, 55–71 (2018)
31. Zhou, X., Nie, X., Li, Z., Lin, X., Xue, E., Wang, L., Lan, J., Chen, G., Du, M., Tong, T.: H-net: A dual-decoder enhanced fcnn for automated biomedical image diagnosis. Information Sciences **613**, 575–590 (2022)