This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Cycle Context Verification for In-Context Medical Image Segmentation

Shishuai Hu[1,2⋆], Zehui Liao[1,2⋆], Liangli Zhen[2], Huazhu Fu[2(✉)], and Yong Xia[1,3,4(✉)]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore 138632, Singapore
[3] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
[4] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China
hzfu@ieee.org; yxia@nwpu.edu.cn

Abstract. In-context learning (ICL) is emerging as a promising technique for achieving universal medical image segmentation, where a variety of objects of interest across imaging modalities can be segmented using a single model. Nevertheless, its performance is highly sensitive to the alignment between the query image and in-context image-mask pairs. In a clinical scenario, the scarcity of annotated medical images makes it challenging to select optimal in-context pairs, and fine-tuning foundation ICL models on contextual data is infeasible due to computational costs and the risk of catastrophic forgetting. To address this challenge, we propose Cycle Context Verification (CCV), a novel framework that enhances ICL-based medical image segmentation by enabling self-verification of predictions and accordingly enhancing contextual alignment. Specifically, CCV employs a cyclic pipeline in which the model initially generates a segmentation mask for the query image. Subsequently, the roles of the query and an in-context pair are swapped, allowing the model to validate its prediction by predicting the mask of the original in-context image. The accuracy of this secondary prediction serves as an implicit measure of the initial query segmentation. A query-specific prompt is introduced to alter the query image and updated to improve the measure, thereby enhancing the alignment between the query and in-context pairs. We evaluated CCV on seven medical image segmentation datasets using two ICL foundation models, demonstrating its superiority over existing methods. Our results highlight CCV's ability to enhance ICL-based segmentation, making it a robust solution for universal medical image segmentation. The code will be available at https://github.com/ShishuaiHu/CCV.

Keywords: Medical image segmentation · In-context learning · Test-time optimization.

⋆ S. Hu and Z. Liao contributed equally. Corresponding authors: H. Fu and Y. Xia.

1 Introduction

Universal medical image segmentation has garnered significant attention due to the increasing need for medical professionals to segment and analyze various objects of interest across diverse imaging modalities, rather than focusing on a single organ or lesion [12,4,32,10,13,9,5,14,29]. In-context learning (ICL) [7], which enables a single model to segment varied targets through contextual guidance[4], has shown promise in addressing this challenge.

Initially developed for large language models (LLMs), ICL leverages contextual information to improve predictions [7] and has been successfully adapted to both natural [3,26,25] and medical image segmentation [4]. In a typical ICL framework for image segmentation, a test image (query) is processed using in-context image-mask pairs that exemplify the target structure. When trained on diverse, large-scale datasets, these models exhibit universal segmentation capabilities [3,26,25,4]. Despite their potential, the performance of these models is highly sensitive to the alignment between the query image and the in-context pairs [31]. Consequently, many research efforts have been devoted to selecting optimal in-context pairs for each test query [28,8,27,22,21]. However, the ideal context for a given query may not exist within the available in-context pairs. Moreover, in medical image segmentation, particularly when dealing with underrepresented targets, the availability of suitable in-context pairs is often limited [6], making their selection impractical. Furthermore, fine-tuning the ICL model on these in-context pairs is not feasible due to computational costs and the risk of catastrophic forgetting [11]. In this paper, we suggest an alternative approach: enhancing the context rather than selecting optimal pairs or fine-tuning the ICL model to improve segmentation performance.

Current methods for visual context enhancement in ICL are limited. One relevant approach is InMeMo [30], which introduces task-specific prompts added to the borders of the context image to guide the ICL model during segmentation. However, InMeMo requires a substantial amount of labeled data to train these task-specific prompts, and the data must correspond to the same task as the test query image. This is particularly impractical in medical image segmentation due to the scarcity of suitable in-context pairs [6]. To overcome this limitation, we propose learning a query-specific prompt to enhance the context for each query image and optimize it with the limited available in-context pairs during testing. However, defining an optimization objective at test time to fully exploit labeled in-context pairs remains a challenge.

In the field of LLMs, test time scaling is gaining increasing attention [20,15]. This approach incorporates constraints, such as budget forcing [15], to prompt the model to verify its outputs and correct reasoning errors. Inspired by the concept of answer verification, we propose a cycle context verification mechanism (see Fig. 1), which allows the ICL model to double-check its predicted mask and optimize the query-specific prompt to improve segmentation. Specifically, given a test query image and an associated in-context pair, the ICL model generates an initial segmentation mask for the query. We then swap the roles of the query and in-context pair: the query image and its predicted mask become the new context

pair, while the original in-context image becomes the new query image. By taking the swapped query and context pair as input, the ICL model can double-check its original prediction for the query image and generate a mask prediction for the in-context image. The accuracy of the original query prediction indicates the alignment between the query image and the context pair, which also impacts the accuracy of the in-context image prediction. In other words, the accuracy of the in-context prediction implies the accuracy of the original query prediction. Since the in-context mask is available, the accuracy of the in-context prediction can be directly evaluated. This enables optimization of the query-specific prompt, thus improving the accuracy of the in-context prediction and, in turn, enhancing the query segmentation.

In this paper, we introduce Cycle Context Verification (CCV) for in-context medical image segmentation. CCV employs a cyclic pipeline during testing that allows the ICL model to verify its query prediction. A learnable, query-specific prompt is incorporated into the query image to improve the alignment between the query and the in-context pairs. This prompt is iteratively updated to improve the accuracy of the in-context prediction, leading to better query segmentation. We evaluate the proposed method against other techniques using two ICL backbones on seven medical image segmentation datasets, demonstrating its superior performance over existing methods. The contributions of this work are three-fold:

1. We propose a novel, plug-and-play cycle context verification pipeline that enables the ICL model to double-check its query predictions at inference time.
2. Unlike traditional prompt learning, which tunes a unified prompt across the entire dataset, we introduce a learnable, query-specific prompt that is optimized to improve the alignment between each query image and its in-context pairs, thereby enhancing segmentation performance.
3. Our experiments on seven medical image segmentation datasets, using two ICL foundation models, show that the proposed method outperforms existing techniques.

2 Method

2.1 Problem Formulation and Method Overview

Let $\mathcal{M}$ denote a pretrained ICL model for medical image segmentation, and let $S = \{(x_n, y_n)\}_{n=1}^{N}$ represent a set of in-context image-mask pairs, where N is the number of available pairs. S can be the limited available pairs or context pairs selected through popular context selection methods [31,18,8]. Given a test query image x_t, the objective is to enhance the alignment between S and x_t to improve the predicted segmentation mask $\hat{y_t^p} = \mathcal{M}(f(x_t), S)$, while keeping model $\mathcal{M}$ frozen. Here, $f(\cdot)$ represents a function that modifies x_t.

The proposed CCV framework introduces a Cycle ICL pipeline, which enables the model to double-check its query prediction y_t^p iteratively, a proxy accuracy measurement $\mathcal{A}$ to evaluate the accuracy of y_t^p, and a learnable query-specific

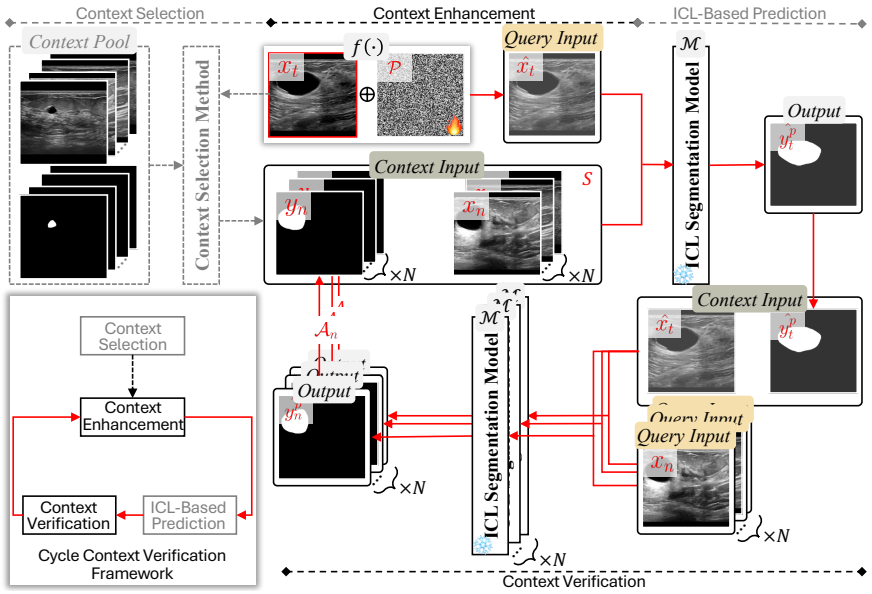

Fig. 1: Diagram of the proposed CCV framework. The symbol $\oplus$ denotes the addition of the query-specific prompt to the query image. The red arrows represent the cycle data flow.

prompt $\mathcal{P}$ to refine x_t towards generating more accurate y_t^p. In the Cycle ICL pipeline, a segmentation mask y_t^p is firstly predicted by $\mathcal{M}$ as usual. The model then swaps the roles of the query and in-context pair, treating the query and its predicted mask as a new in-context pair while designating the original in-context image as the new query. This process yields a secondary prediction y_n^p, whose accuracy, measured using $\mathcal{A}$, serves as an indirect estimate of the reliability of y_t^p. The query-specific prompt $\mathcal{P}$ is iteratively optimized to maximize the accuracy measurement $\mathcal{A}$. Once converged, the enhanced inputs $f(x_t)$ and S are fed to $\mathcal{M}$ to yield the final prediction $\hat{y}_t^p$. An overview of the framework is illustrated in Fig. 1. We now delve into its details.

2.2 Cycle ICL Pipeline for Context Verification

Given a set of in-context pairs $S = \{(x_n, y_n)\}_{n=1}^N$, a pretrained ICL model $\mathcal{M}$, and a query image x_t, the initial query prediction y_t^p is generated as:

$$y_t^p = \mathcal{M}(x_t, S). \tag{1}$$

To enable self-verification, the model's input configuration is modified. The predicted mask y_t^p is paired with the original query image x_t to create a new in-context pair $\{(x_t, y_t^p)\}$, while the original in-context image x_n is designated

as the new query. The model then generates a secondary prediction y_n^p:

$$y_n^p = \mathcal{M}(x_n, \{(x_t, y_t^p)\}). \tag{2}$$

Since the ground truth mask y_n for x_n is available, the accuracy of y_n^p can be quantitatively evaluated using the Dice Similarity Coefficient (DSC) $\mathcal{A}_n = \mathcal{D}(y_n^p, y_n)$. The overall accuracy $\mathcal{A}$ across all in-context pairs is then computed as:

$$\mathcal{A} = \frac{\sum_{n=1}^{N} \mathcal{A}_n}{N}. \tag{3}$$

The segmentation quality of $\mathcal{M}$ is contingent on the context provided. If the initial query prediction y_t^p deviates substantially from the ground truth, incorporating it into the context may degrade the accuracy of y_n^p. Thus, $\mathcal{A}$ serves as an implicit measure of the reliability of y_t^p.

2.3 Query-Specific Prompt for Context Enhancement

To improve segmentation accuracy, the objective is to maximize $\mathcal{A}$. However, directly optimizing $\mathcal{M}$ may potentially lead to model collapse, where the model becomes overfitted to the verification step without improving y_t^p. To circumvent this, we introduce a query-specific prompt $\mathcal{P}$, which is learnable and spatially aligned with the query image x_t. By adding $\mathcal{P}$ to x_t, the correspondence between the query and in-context images is expected to be enhanced, thereby yielding improved segmentation outcomes. The transformation of the query image can be expressed as:

$$\hat{x_t} = f(x_t) = x_t + \mathcal{P}. \tag{4}$$

Using the modified query $\hat{x_t}$, we recompute $\hat{y_t^p}$, $\hat{y_n^p}$, $\hat{\mathcal{A}}_n$, and $\hat{\mathcal{A}}$. The loss function is defined as:

$$\mathcal{L} = 1 - \hat{\mathcal{A}}. \tag{5}$$

Minimizing $\mathcal{L}$ encourages the model to refine $\mathcal{P}$ iteratively. A higher $\hat{\mathcal{A}}$ indicates an improved performance of the predicted masks $\{\hat{y_n^p}\}_{n=1}^{N}$, thereby improving the reliability of y_t^p. To maintain computational efficiency, an accuracy threshold $\mathcal{T}$ is established. If $\hat{\mathcal{A}} > \mathcal{T}$, prompt optimization is halted. Additionally, a maximum iteration limit $\mathcal{E}$ is imposed to prevent excessive optimization. Once $\mathcal{P}$ is optimized, the final segmentation mask $\hat{y_t^p}$ is generated by feeding $\hat{x_t}$ and S into $\mathcal{M}$.

3 Experiments and Analysis

3.1 Experimental Setup

ICL Backbones. The proposed method is evaluated using UniverSeg [4] and SegGPT [26] as ICL backbones. UniverSeg is an ICL backbone specifically designed for medical image segmentation. SegGPT, on the other hand, is an ICL backbone primarily trained on large-scale natural image segmentation datasets.

Table 1: The DSC (%) of the competing methods and those equipped with CCV.

$\mathcal{M}$	Methods	BUS	CQE	GLS	NPP	OCT	ROC	ROD	USC	Average
UniverSeg [4]	RS	47.49	87.21	60.50	28.84	34.97	57.72	77.83	80.79	59.42
	RS+CCV	50.99	87.79	60.71	33.52	38.63	60.26	79.65	81.92	61.68$_{\uparrow 2.26}$
	VPR [31]	54.13	90.68	60.06	32.34	44.72	61.25	78.76	83.91	63.23
	VPR+CCV	57.28	91.11	62.82	36.38	47.73	63.39	80.27	85.22	65.53$_{\uparrow 2.30}$
	DualSC [8]	53.64	88.60	62.29	34.47	46.34	62.10	79.30	84.07	63.85
	DualSC+CCV	57.58	91.11	62.91	36.84	48.03	63.66	80.44	85.11	65.71$_{\uparrow 1.86}$
	InMeMo [30]	48.01	86.73	61.98	28.56	34.53	54.41	79.33	81.27	59.35
	InMeMo+CCV	51.67	87.34	62.15	31.53	37.34	58.01	80.11	82.37	61.32$_{\uparrow 1.97}$
SegGPT [26]	RS	42.05	91.00	76.66	59.89	35.85	67.28	91.38	80.88	68.12
	RS+CCV	46.26	91.79	80.76	64.96	37.67	71.12	92.46	84.13	71.14$_{\uparrow 3.02}$
	VPR [31]	55.92	93.44	85.09	68.29	58.26	68.06	91.91	82.64	75.45
	VPR+CCV	60.71	93.85	86.46	72.04	59.32	72.13	92.59	85.54	77.83$_{\uparrow 2.38}$
	DualSC [8]	57.43	93.55	85.94	68.97	53.54	70.92	92.00	83.23	75.70
	DualSC+CCV	60.79	93.88	86.73	71.58	60.64	72.19	92.62	85.38	77.98$_{\uparrow 2.28}$
	InMeMo [30]	59.68	95.37	81.85	63.50	46.42	78.75	93.22	82.83	75.20
	InMeMo+CCV	61.77	95.57	83.09	66.26	46.85	82.18	94.03	86.20	76.99$_{\uparrow 1.79}$

Datasets and Evaluation Metrics. We selected seven medical image segmentation datasets encompassing eight tasks across seven medical imaging modalities[5] [32] to evaluate the proposed method. These include the BreastUS (BUS, 647 images) [2], COVID-QU-Ex (CQE, 5826 images) [23], GlaS (GLS, 165 images) [19], NeoPolyp (NPP, 1000 images) [16], OCT-CME (OCT, 1460 images) [1], REFUGE (ROC and ROD, 1200 images) [17], and UWaterlooSkinCancer (USC, 206 images) [24] datasets. Each dataset is partitioned into proportions of 3:1:1, designated for test samples, validation samples, and candidate in-context samples, respectively. Notably, these datasets are not included in the training sets of either UniverSeg or SegGPT. The DSC is used as the evaluation metric to assess the segmentation performance.

Implementation Details. The proposed method is implemented using the PyTorch framework and tested on a single NVIDIA 3090 GPU. For the UniverSeg model, the context size is set to eight. All input images are resized to 128 $\times$ 128, converted to gray-scale images, and normalized to the intensity range of [0, 1]. For the SegGPT backbone, the context size is set to one. All images are resized to 448 $\times$ 448, converted to three-channel color images, and normalized by subtracting the mean and dividing by the standard deviation. The query-specific prompt $\mathcal{P}$ is zero-initialized. The threshold $\mathcal{T}$ and the iteration limitation $\mathcal{E}$ are empirically set to 0.9 and 20 respectively.

[5] These datasets can be accessed through https://huggingface.co/datasets/microsoft/BiomedParseData.

3.2 Experimental Results and Analysis

We employ three in-context pair selection methods as baselines to select in-context pairs from the candidate in-context dataset for each test query sample, including: Random Selection (RS), which involves the random selection of in-context pairs from the candidate in-context dataset, VPR [31], which fine-tunes a pretrained CLIP image encoder [18] to extract and rank features, subsequently selecting the top-K in-context pairs, and DualSC [8], which incorporates mask similarity into the ranking process upon VPR. We also incorporate InMeMo [30], a context enhancement method that prompts the model regarding the specific segmentation task to be performed, as a baseline method. The proposed method is designed to be complementary to these existing approaches. Therefore, we integrate the proposed method with these baselines and report the performance.
Comparison using UniverSeg as Backbone. The Dice scores attained by other competing methods, as well as those equipped with the proposed method, are presented in the first part of Table 1. The results indicate that our method improves the overall segmentation performance of RS, VPR, and DualSC, emphasizing the necessity of enhancing context during inference. Furthermore, RS equipped with our method outperforms InMeMo, affirming that query-specific prompts can be superior to task-specific prompts. Our approach also enhances the performance of InMeMo, indicating that the query-specific prompt design is complementary to task-specific prompt.
Comparison using SegGPT as Backbone. The Dice scores for our model and competing methods using SegGPT as backbone are presented in the second part of Table 1. Although using fewer in-context pairs, SegGPT still outperforms UniverSeg, highlighting the strong generalization capabilities of SegGPT. Even though, the proposed method enhances the overall segmentation performance across all competing approaches, further demonstrating its advantages.

Table 2: Performance of removing cycle ICL pipeline and prompt.

Method		Average
Cycle ICL Pipeline	Prompt Optimiz.	
×	×	58.34
✓	×	57.24
×	✓	58.57
✓	✓	60.65

Table 3: Performance of four variants of prompt optimization.

Method				Average
Border Prompt	Image Prompt	Update First	Update Second	
×	✓	✓	×	59.76
×	✓	×	✓	59.37
×	✓	✓	✓	60.65
✓	×	✓	✓	58.82

3.3 Ablation Analysis

To evaluate the effectiveness of the proposed method in a limited context scenario, we select a small number of in-context pairs, *i.e.*, eight for UniverSeg, from the candidate in-context dataset for each task, utilizing them as context to guide the segmentation and evaluate the performance of the ICL model and

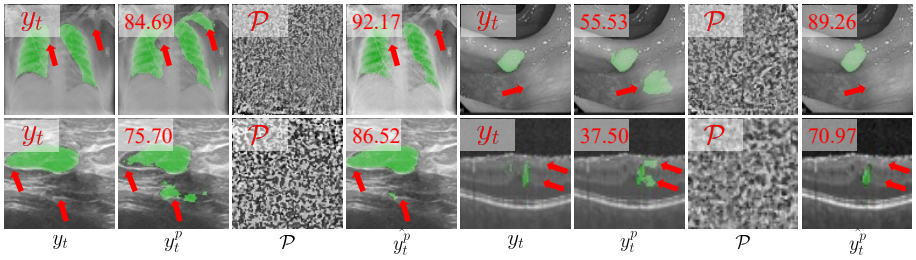

Fig. 2: Visualization of four learned query-specific prompt $\mathcal{P}$, corresponding predictions $\hat{y}_t^p$ and original predictions y_t^p. Together with ground truth y_t. The dice scores of y_t^p and $\hat{y}_t^p$ is also displayed for comparison. Note that $\mathcal{P}$ has been rescaled to [0, 255] for a better visualization. Best viewed in color.

the proposed method. The effectiveness of the cycle ICL pipeline and the query-specific prompt optimization is also evaluated in the limited context scenario via removing these two components respectively. UniverSeg is used as the ICL backbone for the ablation experiments. The results are shown in Table 2. Removing prompt optimization represents employing solely the cycle ICL pipeline and updating the entire ICL model with a small learning rate 1e-4 for 20 iterations. Removing Cycle ICL pipeline represents employing only the query-specific prompt and minimizing the entropy of $\hat{y}_t^p$ to update $\mathcal{P}$. The performance drop of removing prompt optimization can be attributed to the model collapse when updating the entire ICL model. The results reveal that both the cyclical design and the query-specific prompt-based optimization are essential for the functioning of the proposed method.

Analysis of Query-specific Prompt. We compare four configurations in the limited context scenario using the cycle ICL pipeline: use image prompt (CCV), use border prompt [30] instead of image prompt, update the image prompt only in the first stage when generating $\hat{y}_t^p$, and update the image prompt only in the second stage when generating $\hat{y}_n^p$. The results is shown in Table 3. It indicates that the performance of using border prompt is inferior to that of using image prompt. Notably, freezing the image prompt during the generation of either $\hat{y}_t^p$ or $\hat{y}_n^p$ is found to be less effective than updating the image prompt in these two stages. This observation underscores the necessity for the updating of $\mathcal{P}$ to enhance the alignment between the query image and in-context pairs throughout these stages. We visualized examples of the learned query-specific prompt $\mathcal{P}$ in Fig. 2. The predictions generated by the ICL model for x_t and $\hat{x_t}$ are also illustrated. Additionally, the ground truth segmentation maps are provided for reference. The visualization demonstrates that the learned $\mathcal{P}$ can effectively alter x_t towards generating markedly improved segmentation results, thereby underscoring the validity of the proposed methodology.

4 Conclusion

In this paper, we introduced the Cycle Context Verification (CCV) framework to enhance in-context medical image segmentation, particularly under limited in-context data. By implementing a cycle ICL pipeline, the ICL model is capable of double-checking its predictions. The integration and optimization of a query-specific prompt enhance the alignment between the test query image and the associated in-context pairs. Experimental results demonstrate that CCV not only improves existing context selection and enhancement methods, but also elevates segmentation performance in challenging limited context scenarios. Furthermore, ablation studies elucidate the design of the cycle ICL pipeline and query-specific prompt-based optimization. These findings highlight the potential of CCV to advance in-context learning based universal medical image segmentation.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62171377 and 92470101, in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06), in part by the Shenzhen Science and Technology Program under Grant JCYJ20220530161616036, and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grants CX2023016 and CX2022056.

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

References

1. Ahmed, Z., Panhwar, S.Q., Baqai, A., Umrani, F.A., Ahmed, M., Khan, A.: Deep learning based automated detection of intraretinal cystoid fluid. International Journal of Imaging Systems and Technology **32**(3), 902–917 (2022)
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
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. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. Medical Image Analysis **98**, 103310 (2024)
6. Cheng, Z., Wang, S., Xin, T., Zhou, T., Zhang, H., Shao, L.: Few-shot medical image segmentation via generating multiple representative descriptors. IEEE Transactions on Medical Imaging (2024)
7. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al.: A survey on in-context learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1107–1128 (2024)

8. 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 (2024)
9. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
10. 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)
11. Lin, Y., Tan, L., Lin, H., Zheng, Z., Pi, R., Zhang, J., Diao, S., Wang, H., Zhao, H., Yao, Y., et al.: Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. arXiv preprint arXiv:2309.06256 (2023)
12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
13. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023)
14. Mei, J., Zhou, T., Huang, K., Zhang, Y., Zhou, Y., Wu, Y., Fu, H.: A survey on deep learning for polyp segmentation: Techniques, challenges and future trends. Visual Intelligence **3**(1), 1 (2025)
15. Muennighoff, N., Yang, Z., Shi, W., Li, X.L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., Hashimoto, T.: s1: Simple test-time scaling. arXiv preprint arXiv:2501.19393 (2025)
16. 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: Advances in Visual Computing: 16th International Symposium, ISVC 2021, virtual event, October 4-6, 2021, proceedings, part II. pp. 15–28. Springer (2021)
17. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Analysis **59**, 101570 (2020)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
19. Sirinukunwattana, K., Snead, D.R., Rajpoot, N.M.: A stochastic polygons model for glandular structures in colon histology images. IEEE Transactions on Medical Imaging **34**(11), 2366–2378 (2015)
20. Snell, C., Lee, J., Xu, K., Kumar, A.: Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314 (2024)
21. Sun, Y., Chen, Q., Wang, J., Wang, J., Li, Z.: Exploring effective factors for improving visual in-context learning. arXiv preprint arXiv:2304.04748 (2023)
22. Suo, W., Lai, L., Sun, M., Zhang, H., Wang, P., Zhang, Y.: Rethinking and improving visual prompt selection for in-context learning segmentation. In: European Conference on Computer Vision. pp. 18–35. Springer (2024)
23. Tahir, A.M., Chowdhury, M.E.H., Qiblawey, Y., Khandakar, A., Rahman, T., Kiranyaz, S., Khurshid, U., Ibtehaz, N., Mahmud, S., Ezeddin, M.: Covid-qu-ex dataset. Kaggle (2021), https://doi.org/10.34740/kaggle/dsv/3122958

24. Venugopal, V., Joseph, J., Das, M.V., Nath, M.K.: Dtp-net: A convolutional neural network model to predict threshold for localizing the lesions on dermatological macro-images. Computers in Biology and Medicine **148**, 105852 (2022)
25. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6830–6839 (2023)
26. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1130–1140 (2023)
27. 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)
28. Xu, C., Liu, C., Wang, Y., Yao, Y., Fu, Y.: Towards global optimal visual in-context learning prompt selection. arXiv preprint arXiv:2405.15279 (2024)
29. Zeng, Q., Xie, Y., Lu, Z., Xia, Y.: A human-in-the-loop method for pulmonary nodule detection in ct scans. Visual Intelligence **2**(1), 19 (2024)
30. Zhang, J., Wang, B., Li, L., Nakashima, Y., Nagahara, H.: Instruct me more! random prompting for visual in-context learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2597–2606 (2024)
31. Zhang, Y., Zhou, K., Liu, Z.: What makes good examples for visual in-context learning? Advances in Neural Information Processing Systems **36**, 17773–17794 (2023)
32. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., et al.: A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nature Methods pp. 1–11 (2024)