

Unified Medical Lesion Segmentation via Self-referring Indicator

Shijie Chang Xiaoqi Zhao Lihe Zhang* Tiancheng Wang
 Dalian University of Technology

Abstract

The recently emerged in-context-learning-based (ICL-based) models have the potential towards the unification of medical lesion segmentation. However, due to their cross-fusion designs, existing ICL-based unified segmentation models fail to accurately localize lesions with low-matched reference sets. Considering that the query itself can be regarded as a high-matched reference, which better indicates the target, we design a self-referencing mechanism that adaptively extracts self-referring indicator vectors from the query based on coarse predictions, thus effectively overcoming the negative impact caused by low-match reference sets. To further facilitate the self-referencing mechanism, we introduce reference indicator generation to efficiently extract reference information for coarse predictions instead of using cross-fusion modules, which heavily rely on reference sets. Our designs successfully address the challenges of applying ICL to unified medical lesion segmentation, forming a novel framework named SR-ICL. Our method achieves state-of-the-art results on 8 medical lesion segmentation tasks with only 4 image-mask pairs as reference. Notably, SR-ICL still accomplishes remarkable performance even when using weak reference annotations such as boxes and points, and maintains fixed and low memory consumption even if more tasks are combined. We hope that SR-ICL can provide new insights for the clinical application of medical lesion segmentation.

1. Introduction

Medical lesion segmentation (MLS) is an important subtask in the field of medical image analysis with high research value and rich clinical applications [33]. MLS tasks encompass various medical imaging modalities, such as computed tomography (CT), magnetic resonance (MR), and optical coherence tomography (OCT), and various lesions in different body regions, such as colon polyps, thyroid nodules, and brain tumors. Automatic and accurate segmentation of lesions in medical images is an important means of assisting clinical medical diagnosis. In recent years, deep learn-

*Corresponding author.

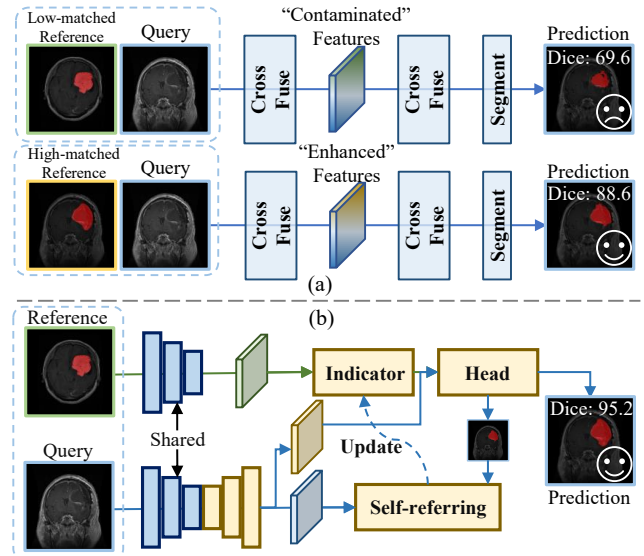


Figure 1. Illustration of different ICL-based unified segmentation frameworks. (a) In current ICL-based models, different reference sets lead to contrasting results due to cross-fusion mechanisms. High-matched reference enhances query features, while low-matched reference contaminates them. (b) Our solution: Individually extract information from reference sets and queries, adaptively extracting information from the query itself for lesion segmentation.

ing methods have become the dominant solution for MLS tasks [14, 34, 48, 59].

Due to the diversity of modalities and lesions, previous methods typically train a separate set of parameters for each modality and each type of lesion to accomplish specific tasks. Traditional segmentation models [6, 21, 35] fail to use a single set of parameters to accomplish various challenging MLS tasks. Inspired by the unified interactive segmentation model SAM [25], MedSAM [29] accomplishes unified interactive medical segmentation based on large-scale hybrid medical datasets. However, the interactive model requires weak annotation of each query image, making it difficult to apply in practice. Recently, many in-context-learning-based (ICL-based) segmentation models have emerged. Benefiting from the ICL paradigm [30],

these generalist models [3, 49, 50] learn unified representations across a large number of diverse tasks, thereby acquiring few-shot transfer capabilities, *i.e.*, reference sets with the same category can indicate the target. UniverSeg [5] and One-Prompt [54] successfully adapt ICL to unified medical segmentation, demonstrating few-shot segmentation capabilities.

However, existing ICL-based methods face two conjugate challenges: 1) Low-matched reference sets lead to erroneous predictions. 2) The cross-fusion designs of reference and query features result in heavy reliance on reference sets, exacerbating the first problem. As shown in Fig. 1 (a), query features are “contaminated” after being cross-fused with low-matched reference, *i.e.*, noises are introduced to query features, while “enhanced” with high-matched reference, leading to significantly contrasting results. Reference selection strategies [18, 42, 58] address the first challenge while seeking high-matched reference sets among numerous images brings additional computational consumption, neglecting the second challenge. In contrast, considering that the query itself can be seen as a high-matched reference without annotation, we can efficiently construct a high-matched image-mask pair by combining the query image itself and coarse prediction extracted based on reference, as shown in Fig. 1 (b). Inspired by this insight, we propose Self-referring Indicator Generation (SRIG), which first uses the cross-attention mechanism to adaptively extract information based on the query feature and rough coarse prediction, and then updates the indicator vectors obtained from the reference to accurately segment the target.

Building on the advantages of the self-referring mechanism, we design SR-ICL to accomplish unified MLS. Fig. 1 (b) illustrates the core perspectives. To fully unlock the potential of SRIG and overcome the feature contamination caused by cross-fusion, we first extract reference and query features individually. Secondly, we introduce a Reference Indicator Generation (RIG) to utilize reference information to generate reference indicator vectors that serve as weights of the dynamic lesion segmentation head to obtain coarse predictions instead of designing complex cross-fusion mechanisms. With RIG and dynamic lesion segmentation head, SR-ICL can efficiently extract reference information, and mitigate the negative impact of low-matched reference sets. Then, SRIG refines indicator vectors by adaptively aggregating information from the query itself. After SRIG, the dynamic lesion segmentation head can accurately localize the lesion targets. In addition, to prevent the model from only focusing on tasks with more data and failing on long-tail tasks, we improve the training strategy to balance the model’s learning across different tasks with fixed and low memory requirements. With the above design, SR-ICL becomes a powerful and effective solution for accomplishing unified MLS.

Our contributions are summarized as follows:

- We propose a self-referring mechanism that adaptively extracts information from the query itself to overcome the negative impact caused by low-matched reference sets.
- We design reference indicator generation to efficiently utilize reference information instead of cross-fusion mechanisms which heavily rely on reference sets, facilitating the self-referring mechanism.
- Our designs successfully apply ICL to unified MLS, forming a novel framework named SR-ICL, which achieves state-of-the-art results on 8 MLS tasks. SR-ICL has impressive performance even with weak reference annotations such as boxes and points. Furthermore, it has relatively low memory requirements during training, offering a new solution for unified MLS.

2. Related Work

2.1. Medical Lesion Segmentation

Medical lesion segmentation (MLS) is a critical task in medical image analysis, aiming to identify and segment lesion regions from medical images. MLS involves multiple modalities, including computed tomography (CT), magnetic resonance (MR), optical coherence tomography (OCT), ultrasound, pathology, endoscopy, and dermoscopy. It also covers various lesions such as skin lesions, brain tumors, and colon polyps. Previous works primarily design task-generic [35] or task-specific [41, 62] network architectures to address MLS. U-Net [35] achieves success in multiple medical image segmentation tasks using an encoder-decoder structure with skip connections. nnUNet [22] improved the model architecture and introduced various tricks to build an advanced task-generic model. TransUNet [9] and SwinUNet [6] utilized Transformer [46] to extract better feature representation. In contrast to task-generic models, some efforts design task-specific models to address the specifics of individual tasks [41, 59, 62]. Existing research often trains a separate model for each task, leading to redundancy and suboptimal results. Moreover, previous methods store lesion knowledge in model parameters, neglecting the benefits of reference images. In this work, we propose a unified MSL framework that uses the same set of parameters for multiple MLS tasks.

2.2. Referring Segmentation

Leveraging reference information to address medical lesion segmentation falls within the scope of referring segmentation. Referring segmentation aims to segment the corresponding target in a query image based on reference information such as image-mask pairs [7], additional modalities [60], and natural language descriptions [56]. Referring Expression Segmentation (RES) and Few-shot Segmentation (FSS) are the two most relevant tasks. RES seg-

ments the target in a query image using natural language descriptions. Previous methods [12, 56] have achieved excellent performance by designing novel vision-language fusion modules. FSS [38] aims to segment the corresponding object with unseen categories of the query image using only a few support image-mask pairs. Existing methods [7, 55] mainly focus on designing novel support-query matching modules to exploit implicit knowledge in pre-trained backbones. The task we solve in this work differs from the above widely studied tasks in the following ways: 1) Existing frozen pre-trained models lack medical lesion knowledge. 2) There is no linguistic description to segment lesions. In this work, we design a unified framework that leverages reference information to solve MLS.

2.3. Unified Segmentation via In-context Learning

GPT-3 [30] introduces in-context learning to NLP, enabling NLP models to perform inference on a series of NLP tasks through given prompts. In-context learning maintains the generalization capability for out-of-domain tasks compared to the multi-task paradigm. MAE-VQGAN [3] and Painter [49] introduce visual in-context learning to computer vision, accomplishing various vision tasks through inpainting and masked image modeling, respectively. Seg-GPT [50] designs a random coloring scheme that forces the model to reference in-context information to achieve unified segmentation. UniverSeg [5] and Spider introduce in-context learning to address unified medical image segmentation and context-dependent concept segmentation, respectively. DSC-ICL [18], SCS [42], and VPR [58] present novel methods for selecting in-context examples to improve performance. However, existing methods primarily focus on the feature fusion and interaction between the query and in-context examples, which introduces computational overhead and neglects the negative impact of low-matched in-context examples and the rich information inherent in the query itself. In this work, we propose a self-prompt module to overcome the above problems.

3. Method

In this section, we first describe the problem formulation of the task in Sec. 3.1. Then, we introduce the overall structure of our method in Sec. 3.2. Next, we provide details of the two proposed components in Sec. 3.3 and Sec. 3.4. The training and inference strategies are in Sec. 3.5.

3.1. Problem Formulation

We define the set of MLS tasks as $\mathcal{D} = \{d_i\}_{i=1}^N$, where d represents each independent task. Previous fully-supervised methods learn a set of parameters θ of model $f(\cdot, \theta)$ for each task to map from image I_d to corresponding ground truth M_d . In contrast, our method learns only one set of parameters to address all tasks in \mathcal{D} . Specifically, episode sampling

is used during both the training and testing phases. Each episode includes a query set $\mathcal{Q} = \{I^q, M^q\}$ and reference set $\mathcal{R} = \{\mathcal{R}_k\}_{k=1}^K$ with the same task d , where \mathcal{R}_k comprises reference image I^r and its binary mask M^r . With \mathcal{R} , the model $f(\cdot, \theta)$ learns to map from I^q to M^q . After training, the model performs episode testing for all tasks without optimization.

3.2. Method Overview

Fig. 2 illustrates the overall architecture of the proposed method. Without loss of generality and simplicity, we introduce the proposed SR-ICL framework when the number of reference images is set to 1, *i.e.*, $\mathcal{R} = \{I^r, M^r\}$. Given a reference image I^r and a query image I^q with a spatial size of $H \times W$, we first extract the semantic features of I^r and I^q , as well as the multi-scale feature set of I^q from a weight-shared encoder. We denote the semantic features as F^r and F^q with the spatial size of $\frac{H}{32} \times \frac{W}{32}$, and multi-scale query feature set as $\{F_i^q\}_{i=1}^4$. $\{F_i^q\}_{i=1}^4$ is then fed into a U-shape skip connection decoder for top-down decoding to obtain $F^{seg} \in \mathbb{R}^{c \times \frac{W}{4} \times \frac{W}{4}}$, where c is the number of channels. After that, a reference indicator generation (RIG) is applied to F^r and M^r to generate reference indicator vectors as weights for the dynamic lesion segmentation head (DLSH). We obtain the initial prediction map $\tilde{M}^{initial}$ by employing DLSH on F^{seg} . Finally, L stacked self-referring indicator generation (SRIG) utilizes the prepared features and initial prediction map to generate the new weights for DLSH. In the l -th SRIG, the $\tilde{M}_{l-1}^{initial}$ and F^q are fed into the SRIG to generate self-referring indicator vectors. The weighted sum of self-referring and reference indicator vectors becomes the new dynamic weights for DLSH to obtain the prediction map $\tilde{M}_l^{initial}$. After all stacked SRIG, F^{seg} is fed into the final DLSH to generate the segmentation map \tilde{M}^{final} .

3.3. Reference Indicator Generation and Dynamic Lesion Segmentation Head

Motivation. Previous ICL-based unified segmentation methods design complex feature interaction modules in the encoder-decoder framework, introducing significant redundant computations and neglecting the negative impact of noise from the reference set. CondInst [45] indicates that the coarse information of interest objects can be encoded as dynamic weights of the FCN segmentation head to segment interest objects. In our framework, the reference features can serve as coarse indicators of the query lesions. Motivated by this insight, we develop the reference indicator generation (RIG) and dynamic lesion segmentation head (DLSH) to accurately segment lesions while reducing redundant computations instead of cross-fusion modules.

Reference Indicator Generation. The goal of RIG is to generate dynamic weights for DLSH using reference fea-

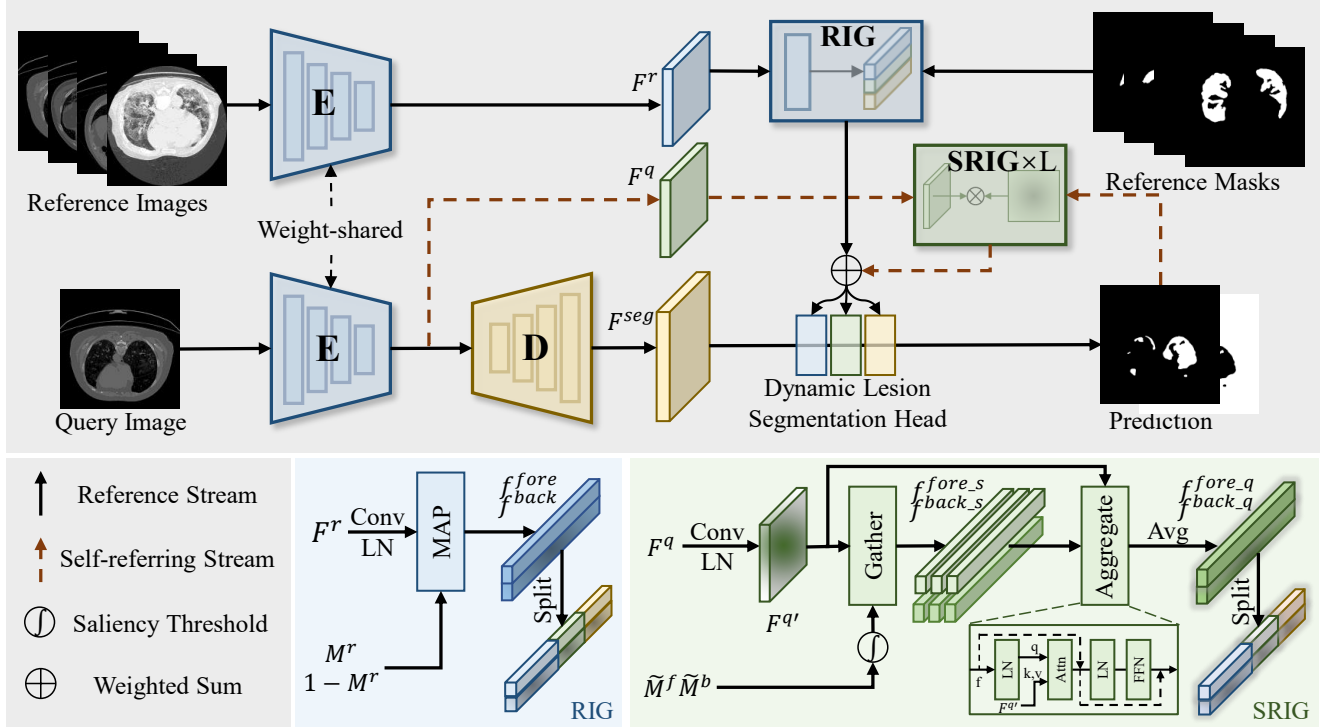


Figure 2. The proposed SR-ICL. It first extracts reference and query features individually. Then, RIG and SRIG generate lesion indicator vectors as weights for dynamic lesion segmentation head to segment target lesions.

ture F^q . To predict the parameters, we first obtain encoded F^{enco} by a learnable convolution, formulated as follows:

$$F^{enco} = \text{LN}(\text{Conv}(F^r)), \quad (1)$$

where Conv and LN denote the 3×3 convolution and layer normalization, respectively. Then, we extract foreground and background indicator vectors through the reference mask as follows:

$$\begin{aligned} f^{fore} &= \text{MAP}(M^r, F^{enco}), \\ f^{back} &= \text{MAP}(1 - M^r, F^{enco}), \end{aligned} \quad (2)$$

where MAP is the masked average pooling operation. After passing through RIG, the dynamic weights required by DLSH are extracted efficiently.

Dynamic Lesion Segmentation Head. DLSH is a simple FCN architecture consisting of 3 stacked 1×1 convolutions with 8, 8, and 1 output channel numbers respectively. Given that the input feature map F^{seg} has c channels, the number of parameters required for DLSH is $c \times 8 + 8 \times 8 + 8 \times 1 = 8c + 72$. In our experiments, c is set to 64, resulting in a total of 584 parameters required. ReLU is used after the first two convolutions as the activation function. We decompose both f^{fore} and f^{back} into three parts, which serve as weights for the corresponding convolution layers, thereby constructing the foreground and background DLSH, respectively. Then,

we obtain the initial foreground and background prediction maps for the query image.

3.4. Self-referring Module

Motivation. Current ICL-based unified segmentation approaches heavily rely on the reference set to segment query images. However, the reference-query matching paradigm is substantially influenced by intra-class pattern variability. Lesions within the same category can exhibit markedly diverse appearance patterns, complicating accurate prediction due to low-matched reference samples. The query image itself can provide supplementary information [17, 26] for lesion segmentation. Inspired by this, we introduce a self-referring mechanism that harnesses the intrinsic information of the query image for lesion segmentation.

Self-referring. The goal of SRIG is to generate self-referring indicator vectors based on the initial prediction maps and the semantic query feature F^q . Given the initial prediction maps \tilde{M}^f and \tilde{M}^b for foreground and background, we filter the prediction maps by saliency threshold τ to obtain salient regions $\tilde{M}^{f'}$ and $\tilde{M}^{b'}$, i.e., setting regions above the τ to 1 and those below to 0. Then, we use a learnable 3×3 convolution and layer normalization to align the channel dimensions of F^q and indicator vectors, formulated as follows:

$$F^{q'} = \text{LN}(\text{Conv}(F^q)), \quad (3)$$

Next, we gather $F^{q'}$ by $\tilde{M}^{f'}$ and $\tilde{M}^{b'}$ to extract salient patches of foreground and background, denoted as f^{fore-s} and f^{back-s} , respectively. The salient patches are insufficient to represent all the information of the query. Therefore, a transformer block with cross-attention is employed to adaptively aggregate query information based on $F^{q'}$, f^{fore-s} , and f^{back-s} , formulated as follows:

$$f^{*-s'} = \text{Transformer}(f^{*-s}, F^{q'}), \quad (4)$$

where f^{*-s} , denoting either f^{fore-s} or f^{back-s} , serves as Q in cross-attention, and $F^{q'}$ functions as K and V . Finally, we average aggregated $f^{fore-s'}$ and $f^{back-s'}$ to obtain self-referring indicator vectors f^{fore-q} and f^{back-q} . We utilize the weighted sum of the reference and self-referring indicator vectors as the updated indicator vectors, formulated as follows:

$$\begin{aligned} f^{fore'} &= \alpha \times f^{fore} + \beta \times f^{fore-q}, \\ f^{back'} &= \alpha \times f^{back} + \beta \times f^{back-q}, \end{aligned} \quad (5)$$

where α and β are hyperparameters. We iteratively apply the self-referring indicator generation L times to progressively update the indicator vectors. After this process, the updated indicator vectors serve as the weights of DLSH to predict the final prediction maps. In our experiments, we observed that a single iteration yields the optimal results. Thus, we set $L = 1$ to balance performance and computational efficiency.

3.5. Training and Inference

Training. Learning unified representations for MLS tasks with diverse modalities is a significant challenge, due to the data imbalance in each task. The model struggles to learn knowledge from long-tail tasks. Spider [61] designs a ‘‘Balance FP - Unify BP’’ strategy to balance all tasks. However, this operation makes the memory required for training relevant to the number of tasks. The resulting high memory requirements prevent training on a large number of tasks. To address this challenge, we adopt ‘‘Separate BP - Average Update’’ during training. In each iteration, we perform forward and backward propagation for each task separately and accumulate the gradients. After processing all tasks, we average the gradients and update the parameters. Our training strategy enables the model to learn more MLS tasks, which has an important role in practical applications. More details are in the supplementary material.

Inference. During inference, we apply the ‘‘argmax’’ function to the concatenation of background and foreground prediction maps to obtain the binary prediction maps. When the number of reference images is greater than 1, we average the indicator vectors generated from each reference image to obtain the weights of DLSH.

Segmentation Task	Dataset	Modality	#Train	#Val
Wet AMD	AMD-SD [20]	OCT	2346	703
Brain Tumor	BTD [10, 11]	MR-T1	2298	766
Adenocarcinoma	EBHI-Seg [39]	Pathology image	636	159
Thyroid Nodule	TNUI 2021 [63]	Ultrasound	966	276
Colon Polyp	Five datasets [4, 23, 40, 43, 47]	Endoscopy image	1450	798
Lung Infection	COVID-19 data [16]	CT	894	383
Breast Lesion	BUSI [1]	Ultrasound	486	161
Skin Lesion	ISIC 2018 [13]	Dermoscopy image	1886	808

Table 1. The dataset information of the 8 selected medical lesion segmentation tasks.

4. Experiment

4.1. Experimental Setup

Datasets. The dataset information of the 8 selected medical lesion segmentation tasks is shown in Tab. 1. AMD-SD [20] annotates 5 pathological manifestations of wet AMD, and we unify these five categories into one category for wet AMD pathological segmentation. BTD [10, 11] divides the dataset into four folds for cross-validation. In our experiments, the first three folds are used for training, and the fourth is used for validation. For EBHI-Seg [39], we divide the subset of Adenocarcinoma segmentation in 4:1 into training and validation sets for experiments. As for the other datasets, we follow previous work [15, 61, 63] to split them.

Evaluation Metrics. Two widely used metrics, Dice score and mean IoU (mIoU), are used for quantitative evaluation. The higher value is better for these metrics.

Implementation Details. All experiments are conducted on a single RTX 3090 GPU with 24G memory and implemented in PyTorch [32]. The model parameters are optimized using Adam [24] scheduled by ‘‘step’’ with an initial learning rate of 0.0001, total steps of 50, decay steps of 30, and decay rate of 0.9. The batch size of reference and query is set to 12 and 4 in each episode, respectively. The spatial resolutions of images are set to 384×384 throughout all experiments. Random flipping and rotating are used as data augmentation techniques. We adopt ConvNeXt-B [28, 53] which can provide strong feature representations as the backbone and use a single set of parameters to accomplish multiple MLS tasks. The sum of structure loss [51] and dice loss [31] between predictions and ground truth is used to train the model.

4.2. Quantitative Evaluation

Setting. We compare our method with specialized, generalist, and unified models. Specialized models refer to models that are trained and tested on a single task. Generalist models refer to models that can be tested directly on the selected 8 MLS tasks without fine-tuning. Unified Models are trained with one set of parameters for the 8 MLS tasks and then tested. Unless otherwise stated, UniverSeg and Spider are tested with 64 randomly sampled image-mask pairs as reference sets, while SegGPT and our SR-ICL are tested

Methods	Methods	Wet AMD		Brain Tumor		Adenocarcinoma		Thyroid Nodule		Polyp		Lung Infection		Breast Lesion		Skin Lesion	
		Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
Specialized Models (One model for one task)																	
TRSRD-Net [2]	ICISN24	-	-	-	-	-	-	85.40	84.65	-	-	-	-	-	-	-	-
LDNet [57]	MICCAI22	-	-	-	-	-	-	-	-	64.25	74.41	-	-	-	-	-	-
WeakPolyp [52]	MICCAI23	-	-	-	-	-	-	-	-	74.90	80.66	-	-	-	-	-	-
Inf-Net [16]	TMI20	-	-	-	-	-	-	-	-	-	-	43.24	52.85	-	-	-	-
DECOR-Net [19]	ISBI23	-	-	-	-	-	-	-	-	-	-	40.25	69.49	-	-	-	-
AAU-Net [8]	TMI22	-	-	-	-	-	-	-	-	-	-	-	-	47.45	65.15	-	-
CMU-Net [44]	ISBI23	-	-	-	-	-	-	-	-	-	-	-	-	54.52	83.02	-	-
MALUNet [36]	BIBM22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	86.32	85.37
EGE-UNet [37]	MICCAI23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	85.88	84.98
UNet [35]	MICCAI15	76.01	80.88	60.44	73.79	91.94	75.33	80.16	83.94	66.29	75.20	54.59	71.10	70.04	75.71	85.93	81.69
TransUNet [9]	MIA24	76.72	81.29	57.92	72.46	91.85	74.23	78.24	82.90	66.41	75.52	43.89	65.33	75.67	80.62	86.40	82.53
RollingUNet [27]	AAAI24	77.13	81.54	61.38	74.71	92.26	75.05	78.96	83.17	69.20	78.33	62.75	76.36	73.84	79.59	87.68	85.04
Generalist Models (One model performs new tasks without fine-tuning)																	
SegGPT [50]	ICCV23	39.42	62.68	24.43	58.61	73.60	61.71	21.33	57.64	60.62	74.05	26.08	58.78	37.79	62.87	29.28	53.62
UniverSeg [5]	CVPR23	45.90	63.66	32.83	61.29	80.85	45.66	60.95	73.63	29.31	53.46	39.98	65.21	65.51	74.16	77.16	76.01
DSC-ICL [18]	TMI24	-	-	-	-	-	-	76.03	-	-	-	-	-	71.88	-	81.51	-
Unified Models (One model for all tasks)																	
UNet [35]	MICCAI15	75.29	80.46	54.95	70.35	91.21	72.66	78.41	82.88	60.31	70.24	47.87	66.66	74.83	79.30	86.88	82.89
TransUNet [9]	MIA24	75.46	80.43	56.59	71.77	90.34	69.16	78.57	83.21	55.81	66.21	36.68	61.62	74.66	79.53	86.14	80.67
RollingUNet [27]	AAAI24	76.46	81.11	60.48	74.43	92.17	71.62	82.26	85.24	62.48	74.53	68.63	79.05	77.58	82.16	87.30	83.92
SegGPT [50]†	ICCV23	71.66	78.64	44.17	67.30	94.24	80.14	79.69	85.18	78.29	84.66	50.10	65.91	78.87	83.87	87.90	85.42
Spider [61]†	ICML24	78.51	82.66	71.59	79.86	93.65	80.30	86.52	88.40	80.66	85.30	73.85	82.07	81.16	84.80	88.79	87.18
Ours	-	80.54	83.84	74.29	81.50	94.96	81.98	87.91	89.18	83.26	86.53	82.36	87.21	84.92	87.08	90.85	87.29

Table 2. Comparison with State-of-the-Art methods in terms of Dice score (%) and mIoU (%). † means these methods are trained or fine-tuned on 8 MLS tasks. The best scores are in **bold**.

using 4 randomly sampled image-mask pairs.

Results. Quantitative comparisons with other state-of-the-art models can be found in Tab. 2. SR-ICL surpasses all other methods and achieves new state-of-the-art results. It can be observed that generalist models [5, 50] with the transfer ability to new tasks perform lower than specialized and unified models. This indicates that existing generalist models are inadequate for clinical demands. In contrast, SR-ICL achieved the best results in all tasks, providing a novel solution for unified MLS.

4.3. Qualitative Evaluation

Setting. In qualitative evaluation, we compare SR-ICL with two unified models SegGPT [50] and Spider [61]. All of these models have been trained or fine-tuned on 8 MLS tasks. The number of selected reference images is set to 1 for better presentation.

Results. We show some visual qualitative comparisons of 8 tasks in Fig. 3. It can be observed that the predictions of our method are closest to the ground truth with only one reference image. SR-ICL can accurately predict lesions of different morphologies, both large and small targets. Only SR-ICL provides accurate predictions for elongated wet AMD lesions (1st row) and small-target lung infections (6th row). With low-matched breast lesion reference (predict large target based on reference with a small target, 7th row), Spider and SegGPT predict normal regions as lesions. We also present some prediction maps of w/o and w/ SRIG in Fig. 4. It can be observed that coarse predictions are refined by

SRIG, demonstrating the effectiveness of SRIG. More qualitative visualizations of SR-ICL with more reference image-mask pairs are in the supplementary material.

4.4. Ablation Study

All ablation results are shown in Tab. 3. Unless otherwise stated, all results are averaged over 5 runs, and the number of randomly selected reference images is set to 1.

Component-wise Ablation. Tab. 3 (a) shows ablation results regarding the effectiveness of proposed components. We construct our baseline by combining a ConvNeXt-B encoder, a U-shape skip connection decoder, and a learnable segmentation head that learns all tasks’ representation. SR-ICL builds upon the baseline by introducing RIG, DLSH, SRIG, and the background stream, while removing the learnable segmentation head. The results show that the proposed RIG and DLSH improve the performance compared to the learnable segmentation head used in the baseline. This indicates that RIG effectively extracts indicator vectors from the reference set. SRIG further improves the performance, suggesting that the extraction of lesion-specific information by the self-referring mechanism helps to accurately localize the lesions. The performance degradation after removing the background data stream shows that the background information helps in lesion segmentation. With the help of the background stream, the model learns the foreground-background feature representation better. We also train SR-ICL separately on each task with the same setting as unified training. It can be observed that the per-

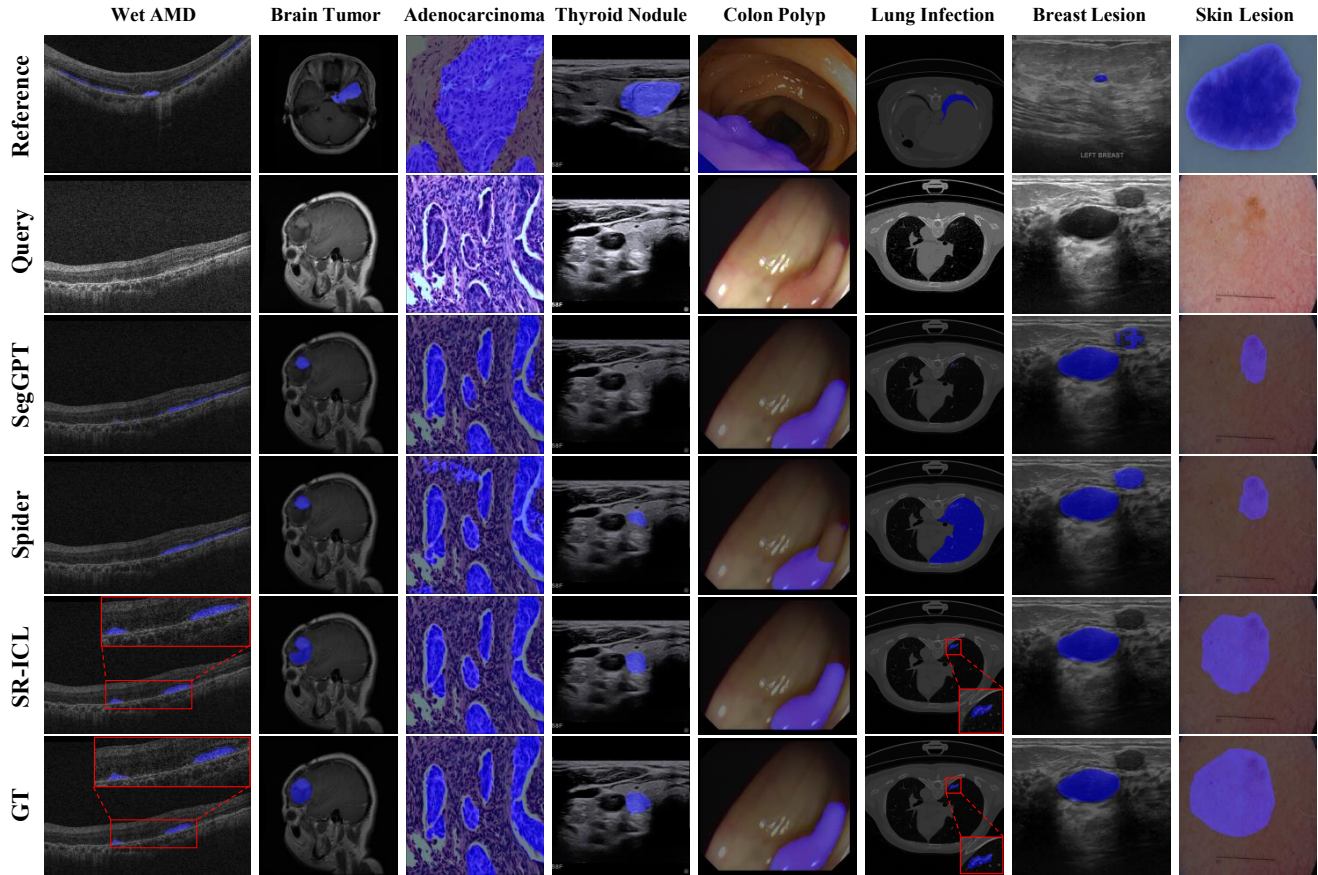


Figure 3. Visualized comparison of SR-ICL, Spider, and SegGPT on 8 diverse MLS tasks. Notably, all models are trained or fine-tuned on 8 MLS tasks. Best viewed on screen.

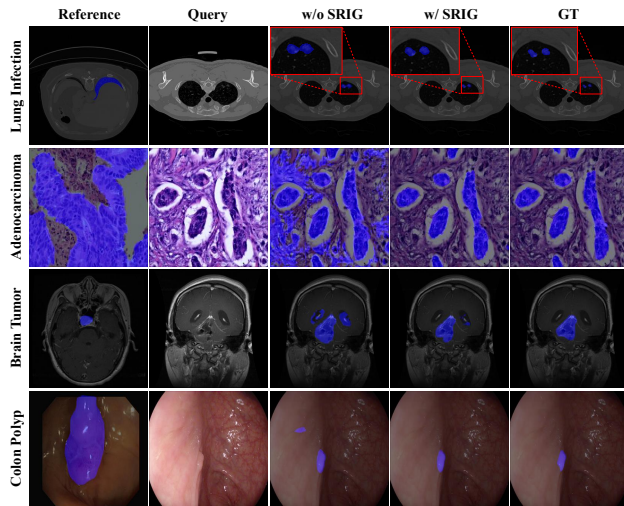


Figure 4. Visualized comparison of w/o SRIG and w/ SRIG.

formance of separate training and unified training is comparable, unlike the results of UNet [35], TransUNet [9], and

RollingUNet [27] in Tab. 2. This indicates that SR-ICL is robust to tasks with diverse modalities.

Hyperparameters of SRIG. SRIG is one of the core components of SR-ICL. We perform ablation experiments on four hyperparameters of SRIG. All results can be found in Tab. 3 (b). The salient regions of the initial prediction map are extracted by saliency threshold τ to control the generation of self-referring indicator vectors. α and β are the weights in the weighted sum of indicator vectors and self-referring indicator vectors, controlling the extent to which SRIG influences the predictions. L denotes the number of iterations of SRIG. The results indicate that the best performance is achieved when $\tau = 0.7$, $\alpha = 0.7$, $\beta = 0.3$, and $L = 2$. Given the small performance difference between $L = 1$ and $L = 2$, we choose $L = 1$ to balance efficiency and performance. The ablation studies on the selection of four hyperparameters demonstrate the robustness of our method to them.

Reference Number. The selection of the reference set is crucial in ICL-based segmentation models. However, benefiting from SRIG, our method maintains performance even

Methods	Wet AMD		Brain Tumor		Adenocarcinoma		Thyroid Nodule		Polyp		Lung Infection		Breast Lesion		Skin Lesion	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
(a) Component-wise Ablation																
Baseline	73.84	78.55	67.34	76.19	90.34	75.70	83.09	84.57	77.83	80.80	74.78	80.52	78.72	81.91	86.02	81.97
w/o SRIG	76.65	79.80	69.52	77.93	92.27	78.00	84.06	85.40	79.42	83.34	77.07	82.35	80.93	83.47	87.55	84.08
w/o Background	77.42	81.40	70.85	78.66	93.10	79.67	85.40	86.39	80.15	84.83	78.41	83.81	81.96	84.35	87.71	84.52
Full (SR-ICL)	79.06	82.79	72.39	80.24	94.30	80.86	86.73	88.26	81.83	85.85	80.07	85.30	83.60	86.09	89.64	86.48
Separate Training	80.70	83.88	70.90	79.26	93.46	80.28	86.52	88.03	81.48	85.70	79.89	84.65	83.06	85.56	89.73	86.63
(b) Hyperparameters of SRIG: τ, α, β, and L																
$\tau = 0.5, \alpha = 0.7, \beta = 0.3, L = 1$	79.01	82.36	72.10	79.84	94.01	80.72	86.50	87.92	81.58	85.67	79.69	84.79	83.66	85.92	89.06	86.20
$\tau = 0.6, \alpha = 0.7, \beta = 0.3, L = 1$	78.47	82.50	72.36	79.94	94.12	80.28	86.58	88.21	81.24	85.64	80.22	85.33	83.10	85.55	89.40	86.20
$\tau = 0.7, \alpha = 0.7, \beta = 0.3, L = 1$	79.06	82.79	72.39	80.24	94.30	80.86	86.73	88.26	81.83	85.85	80.07	85.30	83.60	86.09	89.64	86.48
$\tau = 0.7, \alpha = 0.6, \beta = 0.4, L = 1$	78.58	82.74	72.33	80.15	94.07	80.21	86.14	87.78	81.03	85.28	79.53	85.16	83.13	85.81	89.51	85.97
$\tau = 0.7, \alpha = 0.5, \beta = 0.5, L = 1$	78.34	82.12	72.34	79.44	93.68	80.54	85.99	87.89	80.84	85.95	79.23	84.37	83.62	85.93	89.42	85.92
$\tau = 0.7, \alpha = 0.7, \beta = 0.3, L = 2$	78.83	82.65	72.48	80.27	94.71	81.06	87.18	88.47	81.89	85.93	80.45	85.60	83.79	86.40	89.85	86.60
(c) Number of Randomly Selected Reference Images																
$n = 1$	79.06	82.79	72.39	80.24	94.30	80.86	86.73	88.26	81.83	85.85	80.07	85.30	83.60	86.09	89.64	86.48
$n = 4$	80.54	83.84	74.29	81.50	94.96	81.98	87.91	89.18	83.26	86.53	82.36	87.21	84.92	87.08	90.85	87.29
$n = 16$	80.81	84.21	74.42	81.60	95.26	82.15	88.05	89.19	82.20	86.07	81.94	86.88	84.62	86.99	91.02	87.35
(d) Reference Annotation Type																
Point	79.33	83.03	71.95	79.95	84.12	72.49	86.48	88.05	81.39	85.53	78.18	83.76	82.74	85.60	88.23	85.57
Box	78.40	82.35	71.94	79.95	89.68	77.26	86.31	87.91	81.50	85.58	79.93	85.23	83.28	85.82	89.27	86.03
Mask	79.06	82.79	72.39	80.24	94.30	80.86	86.73	88.26	81.83	85.85	80.07	85.30	83.60	86.09	89.64	86.48

Table 3. Ablation experiments on all tasks in terms of Dice score (%) and mIoU (%), including the effectiveness of components, hyperparameters of SRIG, number of reference examples, and reference annotation type.

Method	Task Number	Batchsize	Memory
Balance FP - Unify BP	8	6/2	16.8G
Balance FP - Unify BP	8	12/4	OOM
Ours	8	6/2	5.7G
Ours	8	12/4	8.3G
Ours	16	12/4	8.3G

Table 4. Ablation studies on training strategies. Batchsize refers to the number of reference images and query images in each iteration. OOM: Out of memory.

with a randomly chosen reference set. We evaluate the performance of SR-ICL with different numbers n of randomly selected reference images, as shown in Tab. 3 (c). In contrast to previous ICL-based methods that set $n = 64$, it can be seen that SR-ICL achieves good results when $n = 4$. SR-ICL is insensitive to n , which benefits from that SRIG efficiently extracts self-referring information.

Reference Annotation Type. We evaluate the effect of different reference annotation types on SR-ICL without fine-tuning. We test the performance of mask, box, and random sampling points over five runs. The number of randomly selected reference images is set to 1. The results are shown in Tab. 3 (d). Except for adenocarcinoma segmentation, the performance degradation is small for all tasks. This demonstrates the generalizability of our approach to different reference annotation types, providing potential for clinical applications. The performance degradation in the adenocarcinoma segmentation task is due to the lesion region occupying a large portion of the image, making it dif-

ficult for the boxes and points to accurately indicate lesion and non-lesion regions.

Training Strategy. As shown in Tab. 4, we validate the effectiveness of the training strategy on a single GPU with 24GB memory. It can be seen that under the same conditions, our method greatly reduces the memory requirement compared to “Balance FP - Unify BP” [61]. Even when doubling the number of tasks, the memory requirement remains the same. Our approach has the advantages of low memory requirements, balancing different tasks, and easy to scale the number of tasks, all towards a more unified MLS.

5. Conclusion

In this paper, we propose SR-ICL with reference and self-referring indicator generation to tackle multiple MLS tasks. The reference indicator generation and dynamic lesion segmentation head alleviate the heavy reliance on reference sets in previous ICL-based methods with cross-fusion modules. Self-referring mechanism overcomes the negative impact caused by low-matched reference by adaptively extracting the query information. We also design a training strategy that can balance numerous MLS tasks with low memory requirements. In our experiments, we find that SR-ICL still achieves impressive performance even with only a few reference images or using weak reference annotations such as boxes and points. These results demonstrate that SR-ICL is an effective solution for MLS in clinical applications, providing new insights for the medical image analysis community.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant 62431004 and 62276046, and by Dalian Science and Technology Innovation Foundation under Grant 2023JJ12GX015.

References

- [1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 5
- [2] Sivadi Balakrishna and Vijender Kumar Solanki. A novel multi-task framework with super-resolution directed network for thyroid nodule segmentation in ultrasound images. In *The International Conference on Intelligent Systems & Networks*, pages 507–515. Springer, 2024. 6
- [3] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *NeurIPS*, 35:25005–25017, 2022. 2, 3
- [4] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilarinho. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 5
- [5] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *ICCV*, pages 21438–21451, 2023. 2, 3, 6
- [6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, pages 205–218. Springer, 2022. 1, 2
- [7] Shijie Chang, Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Beyond mask: Rethinking guidance types in few-shot segmentation. *arXiv preprint arXiv:2407.11503*, 2024. 2, 3
- [8] Gongping Chen, Lei Li, Yu Dai, Jianxun Zhang, and Moi Hoon Yap. Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE TMI*, 42(5):1289–1300, 2022. 6
- [9] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. 2, 6, 7
- [10] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS one*, 10(10): e0140381, 2015. 5
- [11] Jun Cheng, Wei Yang, Meiyang Huang, Wei Huang, Jun Jiang, Yujia Zhou, Ru Yang, Jie Zhao, Yanqiu Feng, Qianjin Feng, et al. Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation. *PLoS one*, 11(6): e0157112, 2016. 5
- [12] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *CVPR*, pages 26573–26583, 2024. 3
- [13] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kallou, Konstantinos Liopyris, Michael Marchetti, 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. 5
- [14] Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek Kumar Singh, Vincent Jaouen, and Dimitris Visvikis. Current and emerging trends in medical image segmentation with deep learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6):545–569, 2023. 1
- [15] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273. Springer, 2020. 5
- [16] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI*, 39(8):2626–2637, 2020. 5, 6
- [17] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *ECCV*, pages 701–719. Springer, 2022. 4
- [18] Jun Gao, Qicheng Lao, Qingbo Kang, Paul Liu, Chenlin Du, Kang Li, and Le Zhang. Boosting your context by dual similarity checkup for in-context learning medical image segmentation. *IEEE TMI*, 2024. 2, 3, 6
- [19] Jiesi Hu, Yanwu Yang, Xutao Guo, Bo Peng, Hua Huang, and Ting Ma. Decor-net: A covid-19 lung infection segmentation network improved by emphasizing low-level features and decorrelating features. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 6
- [20] Yunwei Hu, Yundi Gao, Weihao Gao, Wenbin Luo, Zhongyi Yang, Fen Xiong, Zidan Chen, Yucui Lin, Xinjing Xia, Xiaolong Yin, et al. Amd-sd: An optical coherence tomography image dataset for wet amd lesions segmentation. *Scientific Data*, 11(1):1014, 2024. 5
- [21] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP*, pages 1055–1059. IEEE, 2020. 1
- [22] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. 2
- [23] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020. 5

- [24] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1
- [26] Kurt Koffka. *Principles of Gestalt psychology*. routledge, 2013. 4
- [27] Yutong Liu, Haijiang Zhu, Mengting Liu, Huaiyuan Yu, Zihan Chen, and Jie Gao. Rolling-UNET: Revitalizing MLP’s ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3819–3827, 2024. 6, 7
- [28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 5
- [29] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1
- [30] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1, 2020. 1, 3
- [31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571, 2016. 5
- [32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5
- [33] Dinesh D Patil and Sonal G Deore. Medical image segmentation: a review. *International Journal of Computer Science and Mobile Computing*, 2(1):22–27, 2013. 1
- [34] Imran Qureshi, Junhua Yan, Qaisar Abbas, Kashif Shaheed, Awais Bin Riaz, Abdul Wahid, Muhammad Waseem Jan Khan, and Piotr Szczuko. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90: 316–352, 2023. 1
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1, 2, 6, 7
- [36] Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu. Malunet: A multi-attention and lightweight UNet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1150–1156. IEEE, 2022. 6
- [37] Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-UNET: an efficient group enhanced UNet for skin lesion segmentation. In *MICCAI*, pages 481–490. Springer, 2023. 6
- [38] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, pages 167.1–167.13, 2017. 3
- [39] Liyu Shi, Xiaoyan Li, Weiming Hu, Haoyuan Chen, Jing Chen, Zizhen Fan, Minghe Gao, Yujie Jing, Guotao Lu, Deguo Ma, et al. Ebhi-seg: A novel enteroscopy biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, 10: 1114673, 2023. 5
- [40] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014. 5
- [41] Yongheng Sun, Duwei Dai, Qianni Zhang, Yaqi Wang, Songhua Xu, and Chunfeng Lian. Msca-net: Multi-scale contextual attention network for skin lesion segmentation. *PR*, 139:109524, 2023. 2
- [42] Wei Suo, Lanqing Lai, Mengyang Sun, Hanwang Zhang, Peng Wang, and Yanning Zhang. Rethinking and improving visual prompt selection for in-context learning segmentation. In *ECCV*, pages 18–35. Springer, 2024. 2, 3
- [43] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE TMI*, 35(2):630–644, 2015. 5
- [44] Fenghe Tang, Lingtao Wang, Chunping Ning, Min Xian, and Jianrui Ding. Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, pages 1–5. IEEE, 2023. 6
- [45] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, pages 282–298. Springer, 2020. 3
- [46] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2
- [47] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017. 5
- [48] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET image processing*, 16(5): 1243–1267, 2022. 1
- [49] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pages 6830–6839, 2023. 2, 3
- [50] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *ICCV*, pages 1130–1140, 2023. 2, 3, 6
- [51] Jun Wei, Shuhui Wang, and Qingming Huang. F³net: fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020. 5
- [52] Jun Wei, Yiwen Hu, Shuguang Cui, S Kevin Zhou, and Zhen Li. Weakpolyp: You only look bounding box for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 757–766. Springer, 2023. 6

- [53] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, pages 16133–16142, 2023. [5](#)
- [54] Junde Wu and Min Xu. One-prompt to segment all medical images. In *CVPR*, pages 11302–11312, 2024. [2](#)
- [55] Qianxiong Xu, Guosheng Lin, Chen Change Loy, Cheng Long, Ziyue Li, and Rui Zhao. Eliminating feature ambiguity for few-shot segmentation. In *ECCV*, pages 416–433. Springer, 2024. [3](#)
- [56] Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. In *ECCV*, 2024. [2](#), [3](#)
- [57] Ruifei Zhang, Peiwen Lai, Xiang Wan, De-Jun Fan, Feng Gao, Xiao-Jian Wu, and Guanbin Li. Lesion-aware dynamic kernel for polyp segmentation. In *MICCAI*, pages 99–109. Springer, 2022. [6](#)
- [58] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *NeurIPS*, 36: 17773–17794, 2023. [2](#), [3](#)
- [59] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multi-scale subtraction network. In *MICCAI*, pages 120–130. Springer, 2021. [1](#), [2](#)
- [60] Xiaoqi Zhao, Shijie Chang, Youwei Pang, Jiaying Yang, Lihe Zhang, and Huchuan Lu. Adaptive multi-source predictor for zero-shot video object segmentation. *IJCV*, 132(8):3232–3250, 2024. [2](#)
- [61] Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept segmentation. In *ICML*, 2024. [5](#), [6](#), [8](#)
- [62] Tao Zhou, Yi Zhou, Kelei He, Chen Gong, Jian Yang, Huazhu Fu, and Dinggang Shen. Cross-level feature aggregation network for polyp segmentation. *PR*, 140:109555, 2023. [2](#)
- [63] Xiaogen Zhou, Xingqing Nie, Zhiqiang Li, Xingtao Lin, Ensheng Xue, Luoyan Wang, Junlin Lan, Gang Chen, Min Du, and Tong Tong. H-net: a dual-decoder enhanced fcn for automated biomedical image diagnosis. *Information Sciences*, 613:575–590, 2022. [5](#)