International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 14 Issue 5 ǁ May 2026 ǁ PP. 82-87

DRPA-SAM: Dynamic-Routed Prompt-Aware Adapters for Medical Image Segmentation

Kaige Wang
[*1]*School of Computer Science and Technology, Tongji University , Shanghai, China*
Corresponding Author: Kaige Wang

Abstract
The Segment Anything Model (SAM) provides a general promptable segmentation paradigm, but direct deployment in medical image segmentation is limited by the domain gap between natural and medical images. Medical images often contain low-contrast tissues, ambiguous organ boundaries, small lesions and modality-specific intensity distributions. This paper presents DRPA-SAM, a Dynamic-Routed Prompt-Aware Adapter framework for parameter-efficient adaptation of SAM to medical image segmentation. The pretrained SAM backbone is frozen, and lightweight adapter modules are inserted into the image encoder. Unlike static adapters that use the same transformation for all inputs, the proposed adapter explicitly conditions feature adaptation on prompt embeddings and dynamically routes each sample to multiple medical adapter experts. The dynamic experts learn complementary adaptation patterns such as boundary enhancement, small-object response and low-contrast compensation. Experiments on REFUGE and BTCV demonstrate that DRPA-SAM effectively narrows the domain gap while training only a small number of additional parameters. On BTCV, the proposed method obtains an average Dice of 0.898 with bounding-box prompts, outperforming original SAM, MedSAM and representative medical segmentation networks. These results show that prompt-aware dynamic routing is an effective and efficient strategy for adapting foundation segmentation models to medical images.

Keywords: *Medical image segmentation, Segment Anything Model, adapter tuning, prompt-aware adaptation, dynamic routing, parameter-efficient fine-tuning.*

Date of Submission: 02-05-2026 Date of Acceptance: 13-05-2026

I. INTRODUCTION

Medical image segmentation is a fundamental task in computer-aided diagnosis and treatment planning. It aims to identify anatomical structures or lesions from medical images and directly affects downstream clinical procedures such as lesion measurement, radiotherapy contouring, preoperative planning and follow-up assessment. Although deep segmentation networks have achieved strong performance on public benchmarks, they usually require task-specific training and dense pixel- or voxel-level annotations. In real clinical scenarios, the cost of expert annotation and the distribution shift across scanners, centers and imaging protocols remain major barriers to robust deployment.

SAM has recently attracted broad attention because of its general promptable segmentation ability. By taking points, boxes or masks as prompts, SAM can generate masks for diverse objects without task-specific retraining. However, the model is mainly trained on large-scale natural images. Medical images differ substantially from natural images in imaging mechanism, intensity distribution, texture pattern and semantic structure. Directly applying SAM to CT, fundus or other medical modalities may lead to missed small structures, inaccurate boundaries and unstable predictions. Full fine-tuning on medical data can improve performance but requires high computational cost and may weaken the general representation learned by SAM.

Parameter-efficient fine-tuning provides a more practical solution. Adapter-based tuning freezes the pretrained backbone and learns small modules inserted into the network. Nevertheless, conventional static adapters are not fully suited to promptable medical segmentation. The segmentation behavior of SAM is inherently conditioned on prompts, and different prompt types impose different adaptation requirements. A point prompt may require local feature propagation, a box prompt may provide global spatial constraints, and a mask prompt may carry shape information. Thus, medical-domain adaptation should be aware of prompt information rather than applying a fixed feature transformation to all samples.

To address this problem, this work proposes DRPA-SAM, a Dynamic-Routed Prompt-Aware Adapter method. The core idea is to make the adapter conditioned on prompt embeddings and to route each input to a set of medical adapter experts. The proposed design keeps the SAM backbone frozen, inserts adapters into the ViT image encoder, and uses a lightweight router to combine expert outputs based on image and prompt representations. The method is evaluated on REFUGE and BTCV, covering both fundus image segmentation

and abdominal CT multi-organ segmentation. The experimental results demonstrate strong accuracy and parameter efficiency.

II. RELATED WORK

2.1 Medical Image Segmentation

Deep learning has become the dominant paradigm for medical image segmentation. U-Net[1] and its variants use encoder-decoder structures with skip connections to recover fine spatial details and remain strong baselines for many medical tasks. Later methods introduce attention mechanisms, nested skip connections, three-dimensional convolution and transformer blocks to improve long-range dependency modeling and volumetric context aggregation. Representative models such as nnUNet[2], TransUNet[3], UNETR[4] and Swin-UNetr[5] have achieved competitive performance on multi-organ and lesion segmentation benchmarks. However, these models are usually trained for specific datasets or modalities and may require careful reconfiguration when transferred to new tasks.

2.2 Promptable Foundation Segmentation Models

SAM[6] formulates segmentation as a prompt-conditioned mask generation problem. It contains an image encoder, a prompt encoder and a mask decoder, and can respond to sparse prompts such as points and boxes. This flexible interface is attractive for medical images because sparse clinical interaction is often easier than dense annotation. Medical adaptations of SAM, such as MedSAM[7] and SAM-Med3D[8], improve the model through retraining or domain-specific data. However, full model adaptation is computationally expensive. Moreover, direct prompt decoding may not sufficiently correct medical-domain feature mismatch in the image encoder.

2.3 Parameter-Efficient Adapter Tuning

Adapter tuning inserts small trainable bottleneck modules into a frozen pretrained network.[9] It has been widely used for efficient transfer because it preserves the backbone representation while reducing the number of trainable parameters. In vision transformers, adapters are commonly placed after attention blocks or feed-forward networks. For medical SAM adaptation, static adapters can reduce training cost, but they ignore the fact that SAM predictions are controlled by prompts. DRPA-SAM extends adapter tuning by making the adapter prompt-aware and dynamically routed, thereby aligning the adaptation mechanism with the promptable nature of SAM.

III. METHODOLOGY

3.1 Overview

The proposed framework keeps the pretrained SAM image encoder, prompt encoder and mask decoder mostly unchanged. Trainable adapter modules are inserted into selected transformer layers of the SAM image encoder. During training, the model receives medical images and standard spatial prompts such as points or bounding boxes generated from the available annotations. The prompt encoder converts prompts into embeddings, which are used not only by the mask decoder but also by the proposed adapters to guide image feature adaptation.

3.2 Standard Adapter for SAM Image Encoder

Let X in $R^{N \times C}$ denote the visual token sequence in a transformer block, where N is the number of tokens and C is the channel dimension. A standard adapter uses a bottleneck structure consisting of a down-projection, a nonlinear activation and an up-projection. The output is added back to the original feature through a residual connection. In this work, adapters are inserted into the SAM ViT encoder after multi-head self-attention and along the MLP residual path. Since the SAM backbone is frozen, only the adapter parameters and routing parameters are optimized. Formally, the standard bottleneck adapter is written as:

$$Z = \sigma(XW_{down} + b_{down}), \quad A_{std}(X) = ZW_{up} + b_{up}, \quad X_{out} = X + s \cdot A_{std}(X),$$

where $W_{down} \in R^{C \times d}$ and $W_{up} \in R^{d \times C}$ are the down- and up-projection matrices, d is the bottleneck dimension, $\sigma(\cdot)$ denotes the ReLU activation, and s is a scaling coefficient controlling the strength of the residual adaptation.

3.3 Prompt-Aware Adapter

Static adapters apply the same transformation regardless of the prompt. This is inconsistent with SAM because the desired mask is defined jointly by the image and the prompt. DRPA-SAM therefore introduces prompt-aware modulation into the adapter. Given a prompt embedding P from the SAM prompt encoder, a

lightweight projection network maps P into a modulation vector. This vector recalibrates the bottleneck feature after down-projection. The adapter output is then projected back to the original dimension and added to the visual feature. In this way, the image encoder can produce different adapted features for point prompts, box prompts or mask prompts.

For prompt-aware modulation, the prompt embedding P is first projected into the adapter bottleneck space:

$$m = \varphi(P), \quad Z_p = Z \odot (1 + m), \quad A_{pa}(X, P) = \sigma(Z_p)W_{up} + b_{up}, \quad Xout = X + s \cdot A_{pa}(X, P),$$

where $\varphi(\cdot)$ is a lightweight projection network and $\odot$ denotes element-wise multiplication. This formulation makes the encoder-side feature adaptation explicitly dependent on the current segmentation prompt.

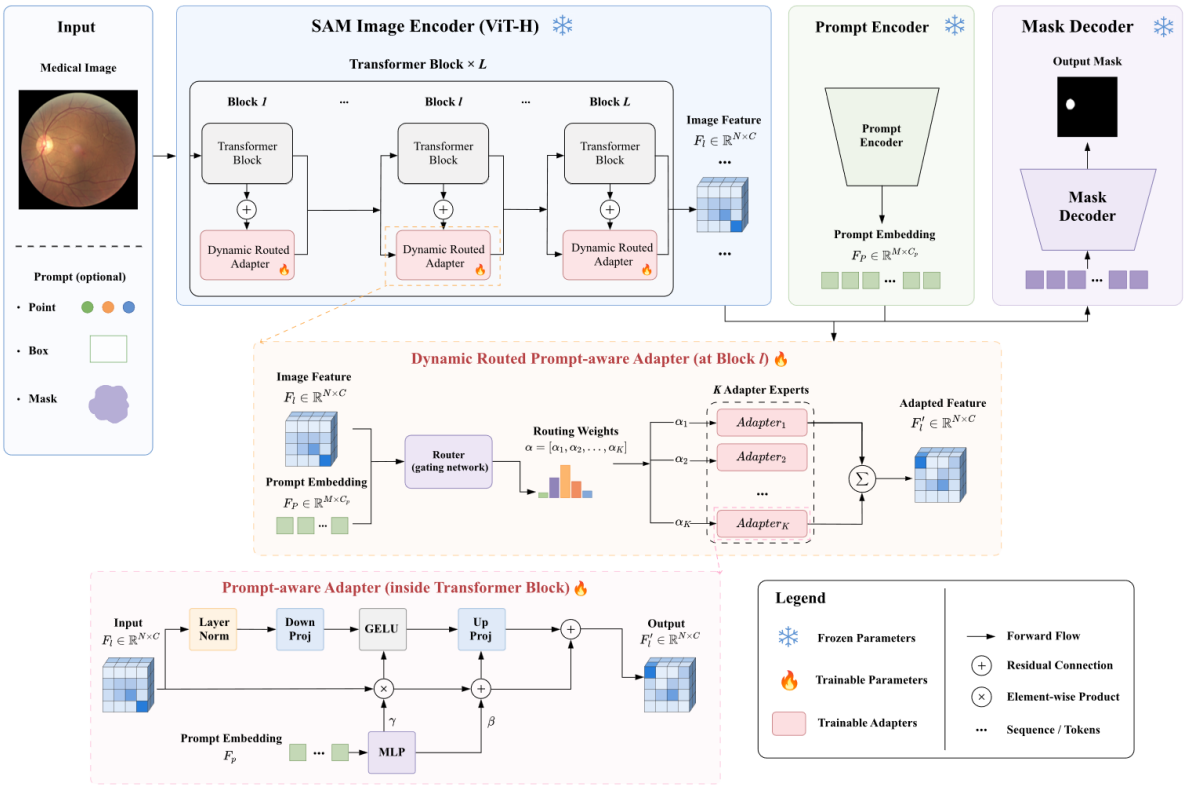

Figure 1: Overall framework of the proposed DRPA-SAM adapter structure

3.4 Dynamic-Routed Medical Adapter Experts

Medical segmentation tasks are heterogeneous. Large organs, small organs, low-contrast tissues and boundary-ambiguous lesions require different feature compensation patterns. To improve conditional modeling, DRPA-SAM replaces a single adapter with multiple prompt-aware adapter experts. A router takes the concatenation of the global image feature and the prompt representation as input and predicts expert weights. The final adaptation output is the weighted sum of all expert outputs. Different experts can specialize in complementary behaviors, such as enhancing boundaries, recovering small structures and correcting modality-specific feature shifts. The router enables sample-adaptive expert selection with limited additional parameters.

Let$\{A_i\}_{i=1}^{K}$ denote K prompt-aware adapter experts. The routing weights are computed from the global image descriptor g(X) and prompt embedding P:

$$\alpha = \text{softmax}(R([g(X); P])), \quad A_{drpa}(X, P) = \Sigma_{i=1}^{K} \alpha_i A_i(X, P), \quad X_{out} = X + s \cdot A_{drpa}(X, P),$$

where $R(\cdot)$ is the router network, $[\cdot;\cdot]$ denotes feature concatenation, and α_i is the contribution of the i-th expert. The softmax normalization encourages a sample-specific but stable expert mixture.

3.5 Training Objective

The model is trained with standard supervised segmentation losses using available pixel-level labels. The loss combines Dice loss and binary cross-entropy loss to account for class imbalance and pixel-wise prediction accuracy. During optimization, the SAM backbone remains frozen, while the prompt-aware adapters and the router are updated. This setting isolates the effect of SAM Adapter adaptation and keeps the study focused on encoder-side parameter-efficient adaptation.

The supervised training objective is defined as

$$L_{seg} = \lambda_{dice} L_{dice}(\hat{Y}, Y) + \lambda_{bce} L_{bce}(\hat{Y}, Y),$$

where $\hat{Y}$ is the predicted mask, Y is the ground-truth mask, and λdice and λbce balance region-level overlap and pixel-level classification terms.

IV. EXPERIMENTAL SETUP

REFUGE is used for fundus optic disc and optic cup segmentation. This dataset evaluates segmentation performance on targets with different scales, especially the smaller and more boundary-sensitive optic cup. BTCV is used for abdominal CT multi-organ segmentation and includes spleen, kidneys, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava and pancreas. The experiments compare DRPA-SAM with original SAM, MedSAM and representative medical segmentation networks, including nnUNet, TransUNet and Swin-UNetr. Dice and IoU are used as the main evaluation metrics. Prompt settings include one point, three points and bounding boxes.

V. RESULTS AND DISCUSSION

5.1 Experimental Results

Table 1 reports REFUGE results. DRPA-SAM obtains consistently better performance than MedSAM under different prompt settings. With BBox 0.75 prompts, it reaches 98.3 Dice for optic disc and 87.5 Dice for optic cup. The improvement on optic cup is important because cup segmentation is more sensitive to small-scale structures and subtle boundaries. This indicates that prompt-aware feature adaptation can better preserve fine medical structures than static or fully generic SAM features.

Table 1: REFUGE segmentation results under different prompt settings.

Method	Param(M)	Trainable(M)	Disc Dice	Disc IoU	Cup Dice	Cup IoU
nnUNet	16	16	94.7	87.3	84.9	75.1
TransUNet	96	96	95.0	87.7	85.6	75.9
Swin-UNetr	138	138	95.3	87.9	84.3	74.5
MedSAM 1 point	636	636	92.9	85.5	82.1	73.8
MedSAM 3 points	636	636	93.8	86.2	82.8	74.2
MedSAM BBox 0.75	636	636	94.6	86.7	82.8	75.9
DRPA-SAM 1 point	636	13	97.4	89.5	86.8	78.8
DRPA-SAM 3 points	636	13	97.9	89.8	87.1	79.0
DRPA-SAM BBox 0.75	636	13	**98.3**	**90.1**	**87.5**	**79.9**

Table 2 shows BTCV results. Original SAM is much weaker than medical segmentation models on abdominal CT because its natural-image representation cannot reliably describe low-contrast organs and complex anatomical shapes. MedSAM improves the results through medical adaptation, but DRPA-SAM achieves the best average Dice. With BBox 0.75 prompts, DRPA-SAM reaches 0.898 average Dice, outperforming Swin-UNetr and MedSAM. The gains on gallbladder, inferior vena cava and pancreas suggest that dynamic routing is especially helpful for challenging small or ambiguous structures.

The qualitative result in Figure 2 further illustrates the advantage of adapter-based SAM adaptation. Compared with original SAM and other baselines, DRPA-SAM better preserves organ completeness and produces smoother boundaries. For organs with weak contrast or small size, such as gallbladder, esophagus and pancreas, the predictions are less fragmented and more consistent with anatomical structures.

The parameter efficiency of DRPA-SAM is another important advantage. Compared with MedSAM, which updates the full SAM model, DRPA-SAM trains only a small number of adapter and router parameters while keeping the 636M-parameter backbone frozen. This makes the approach more practical for medical institutions with limited computational resources and reduces the risk of overfitting on small medical datasets.

Table 2: BTCV per-organ Dice comparison.

Model	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Panc.	Avg
TransUNet	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.775	0.838
Swin-UNetr	0.971	0.936	0.943	0.794	0.773	0.975	0.921	0.892	0.853	**0.794**	0.869
nnUNet	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.680	0.802
SAM 3 points	0.622	0.710	0.812	0.614	0.605	0.513	0.673	0.645	0.483	0.564	0.631
SAM BBox 0.75	0.415	0.621	0.678	0.580	0.595	0.469	0.521	0.612	0.539	0.588	0.550
MedSAM 3 points	0.758	0.831	0.889	0.782	0.733	0.917	0.858	0.876	0.755	0.763	0.820
MedSAM BBox 0.75	0.746	0.842	0.873	0.772	0.745	0.897	0.860	0.889	0.743	0.739	0.804
DRPA-SAM 1 point	0.978	0.935	0.966	0.823	0.818	0.981	0.931	0.915	0.877	0.767	0.883
DRPA-SAM 3 points	0.980	0.936	0.968	0.826	**0.821**	**0.986**	0.934	0.917	0.878	0.771	0.887
DRPA-SAM BBox 0.75	**0.985**	**0.947**	**0.975**	**0.842**	0.808	0.983	**0.942**	**0.939**	**0.899**	0.790	**0.898**

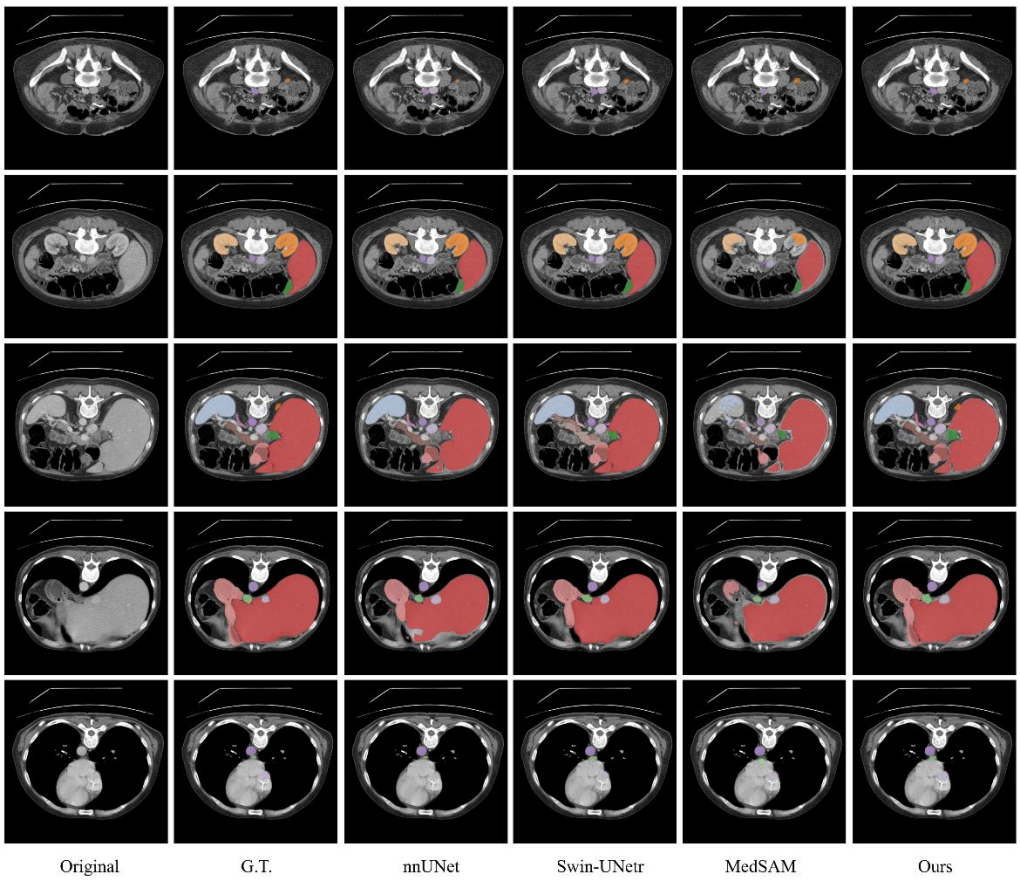

Figure 2: Qualitative segmentation comparison on BTCV

5.2 Ablation Study

To further analyze the effectiveness of each design in DRPA-SAM, ablation experiments are organized around two questions: whether prompt-aware modulation improves over a static adapter, and whether dynamic routing over multiple experts provides additional benefit. The same backbone, datasets, prompt settings and training protocol are kept for all compared variants. Table 3 and Table 4 report Dice scores on REFUGE Disc, REFUGE Cup and BTCV.

Table 3 summarizes the component ablation. The static Adapter provides a strong parameter-efficient medical adaptation baseline. Adding prompt-aware modulation improves the Dice scores on all three targets, and introducing dynamic routing further increases performance. The full DRPA-SAM variant achieves the best results, reaching 0.983 on REFUGE Disc, 0.878 on REFUGE Cup and 0.898 on BTCV.

Table 3: Component ablation of DRPA-SAM

Method	REFUGE Disc	REFUGE Cup	BTCV
Adapter	0.974	0.868	0.883
+ Prompt	0.978	0.872	0.889
+ Routing	0.979	0.873	0.891
All	0.983	0.878	0.898

Table 4 analyzes the number of routed Adapter experts. When the number of experts is small, the model cannot sufficiently represent the feature differences among organs and modalities. Increasing the number to three consistently improves performance. Using four experts does not bring further gains and slightly reduces REFUGE Cup and BTCV results, suggesting that excessive experts introduce extra parameter cost and optimization instability.

Table 4: Ablation on the number of routed adapter experts

Number	REFUGE Disc	REFUGE Cup	BTCV
1	0.978	0.872	0.889
2	0.981	0.875	0.894
3	0.983	0.878	0.898
4	0.983	0.877	0.897

Overall, the ablation results support the design motivation of DRPA-SAM. Prompt-aware modulation aligns Adapter tuning with the prompt-conditioned nature of SAM, and dynamic routing introduces sample-adaptive expert selection for heterogeneous medical images. The best performance is obtained when both mechanisms are used with three Adapter experts.

VI. CONCLUSION

This paper presents DRPA-SAM, a Dynamic-Routed Prompt-Aware Adapter method for adapting SAM to medical image segmentation. The method freezes the pretrained SAM backbone and inserts lightweight prompt-aware adapters into the image encoder. By using dynamic routing over multiple medical expert adapters, DRPA-SAM can model heterogeneous adaptation patterns for different organs, modalities and prompt conditions. Experiments on REFUGE and BTCV show that the proposed method outperforms original SAM, MedSAM and several representative medical segmentation networks while training far fewer parameters than full fine-tuning. Future work will further improve expert specialization, analyze routing behavior across modalities and extend the adapter design to three-dimensional medical segmentation.

REFERENCES

[1]. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," MICCAI, 2015.
[2]. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, 2021.
[3]. J. Chen, Y. Lu, Q. Yu, et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv preprint, 2021.
[4]. A. Hatamizadeh, Y. Tang, V. Nath, et al., "UNETR: Transformers for 3D medical image segmentation," WACV, 2022.
[5]. H. Cao, Y. Wang, J. Chen, et al., "Swin-Unet: Unet-like pure Transformer for medical image segmentation," arXiv preprint, 2021.
[6]. A. Kirillov, E. Mintun, N. Ravi, et al., "Segment Anything," ICCV, 2023.
[7]. J. Ma, Y. He, F. Li, et al., "Segment Anything in Medical Images," Nature Communications, 2024.
[8]. J. Huang, Y. Li, J. Tao, et al., "SAM-Med3D," arXiv preprint, 2023.
[9]. N. Houlsby, A. Giurgiu, S. Jastrzebski, et al., "Parameter-efficient transfer learning for NLP," ICML, 2019.
[10]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2021.
[11]. E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-rank adaptation of large language models," ICLR, 2022.
[12]. M. Jia, L. Tang, B. C. Chen, et al., "Visual prompt tuning," ECCV, 2022.
[13]. K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," IJCV, 2022.
[14]. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," CVPR, 2015.
[15]. G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, 2017.