Med-FastSAM: Improving Transfer Efficiency of SAM to Domain-Generalised Medical Image Segmentation

Yuxiang Luo[1*] **Qing Xu**[2*] **Jinglei Feng**[3] **Guangwu Qian**[4†] **Wenting Duan**[5]
[1]Pittsburgh Institute, Sichuan University
[2]School of Computer Science, University of Nottingham Ningbo China
[3]Department of Radiology, North Sichuan Medical College
[4]West China Biomedical Big Data Center, West China Hospital, Sichuan University
[5]School of Computer Science, University of Lincoln
luoyuxiang2004@outlook.com, qing.xu@nottingham.edu.cn,
g_qian@scu.edu.cn, jingleifeng82@gmail.com,
wduan@lincoln.ac.uk

Abstract

Medical image segmentation is a crucial computer vision task in medical image analysis. Recently, the Segment Anything Model (SAM) has made significant advancements in natural image segmentation. Despite current studies indicating the potential of SAM to revolutionise medical image segmentation using parameter-efficient fine-tuning techniques, it still faces three primary challenges. Firstly, these methods still rely on the large vision transformer of SAM, which is computationally expensive. Secondly, the point and box prompt modes of SAM demand manual annotations, which are time-consuming and expensive in medical scenarios and reduce their clinical applicability. Thirdly, SAM leverages large-size patches to predict masks, resulting in the loss of fine-grained details. To address these limitations, in this paper, we propose a fast-transferring architecture for adapting SAM to domain-generalised medical image segmentation, named Med-FastSAM. Specifically, we introduce a lightweight knowledge aggregation encoder that combines the distilled natural image knowledge with learned medical-specific information for producing feature representation. Moreover, we devise a coarse prompt module to automatically generate coarse masks for guiding segmentation decoding. Furthermore, we design a multi-scale feature decoder to produce precise segmentation masks. Eventually, extensive experiments on four benchmark datasets have been conducted to evaluate the proposed model. The result demonstrates that Med-FastSAM outperforms state-of-the-art methods without any manual prompts. Especially, our model shows excellent zero-shot domain generalisation performance by using only 15.45% parameters compared to the standard SAM. The code for our work and more technical details can be found at https://github.com/GalacticHogrider/Med-FastSAM.

1 Introduction

Medical image segmentation aims to delineate disease regions in complex medical imaging accurately, which is vital for various clinical applications, including diagnosis, treatment planning, and surgical navigation [17]. In the last decade, computer-aid automatic segmentation methods have received the most attention from pathologists [25]. Specifically, deep neural networks, such as U-Net [24], have

*Equal contribution
†Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

significantly contributed to the field. It designs an encoder-decoder structure with symmetric skip connections that combine high-resolution features from the contracting path with the up-sampled output. Inspired by the architecture of U-Net, various variants have been introduced to further improve its performance in segmentation [10, 11]. However, such models demonstrate poor generalisation performance due to the limitation of the local receptive field. The Segment Anything Model (SAM) [3] has emerged as a significant advancement in natural image segmentation tasks. SAM leverages a robust vision transformer to achieve state-of-the-art generalization performance, demonstrating its capability to handle complex scenes with high accuracy and efficiency [6]. While beneficial, adapting SAM to medical image segmentation faces significant challenges. Due to a large Vision Transformer (ViT) [5] as the image encoder, SAM requires huge computational costs. Particularly, the ViT-H image encoder in SAM has 632M parameters. Medical image datasets usually contain limited samples that are difficult to support the global fine-tuning of SAM. Although existing medical SAMs [28, 33] utilise parameter-efficient fine-tuning techniques to reduce the parameters during training, they are still based on the large ViT encoder. Secondly, to generate a precise mask, SAM usually requires manual annotations (e.g. point and box) as prompts, which are time-consuming and expensive in medical scenarios as they depend on expert knowledge. Thirdly, the mask decoder of SAM primarily relies on high-level features extracted by the transformer [12]. This can lead to losing fine-grained details, which are crucial for boundary-sensitive medical segmentation tasks.

To address these issues, we propose a novel architecture for adapting SAM to domain-generalised medical image segmentation, named Med-FastSAM. It includes three modules: lightweight knowledge aggregation encoder, coarse prompt encoder and multi-scale feature decoder. Specifically, the Lightweight Knowledge Aggregation Encoder (LKA-Encoder) combines the natural image knowledge, distilled from SAM, with learned medical-specific information to refine feature representation, which reduces model parameters and computational costs. To eliminate the demand for manual prompts, we devise a Coarse Prompt Encoder (CP-Encoder) that automatically generates coarse masks, providing sufficient prompt information without requiring manual annotations. Additionally, we present a Multi-Scale Feature Decoder (MSF-Decoder) to further enhance segmentation accuracy by incorporating with fine-grained details at different scales.

The contributions of our work are summarized as follows:

- We propose the LKA-Encoder to reduce the computational costs of feature extraction. The image encoder leverages a group attention mechanism and a hybrid expert head to effectively aggregate both natural and medical domain-specific knowledge, enhancing the efficiency of transfer learning.
- We introduce the CP-Encoder to eliminate the requirement of laborious annotations for prompts and improve the applicability in clinical applications. By utilising traditional image processing methods, CP-Encoder automatically generates coarse masks as prompts, guiding the segmentation decoding.
- We devise the MSF-Decoder to achieve the prediction of segmentation masks. It employs a multi-scale sampling approach to capture local fine-grained details, improving the precision of the segmentation masks.
- We take LKA-Encoder, CP-Encoder and MSF-Decoder to build our Med-FastSAM. We evaluate the proposed framework on four datasets, and the results demonstrate that Med-FastSAM outperforms state-of-the-art methods without any manual annotations. Notably, our model exhibits superior domain generalization capabilities.

2 Related Work

2.1 Medical Image Segmentation

Medical image segmentation has traditionally relied on methods like Otsu thresholding and the Watershed algorithm, which, despite their early successes, struggle with generalization and robustness across different imaging conditions. The introduction of U-Net [24] marked a significant advancement in the field, with its encoder-decoder architecture and skip connections allowing for effective contextual feature capture and precise segmentation results. However, the performance of U-Net is highly task-dependent, and its generalization to unseen domains remains a notable challenge [15, 18, 16]. To overcome these limitations, several U-Net variants have been developed.

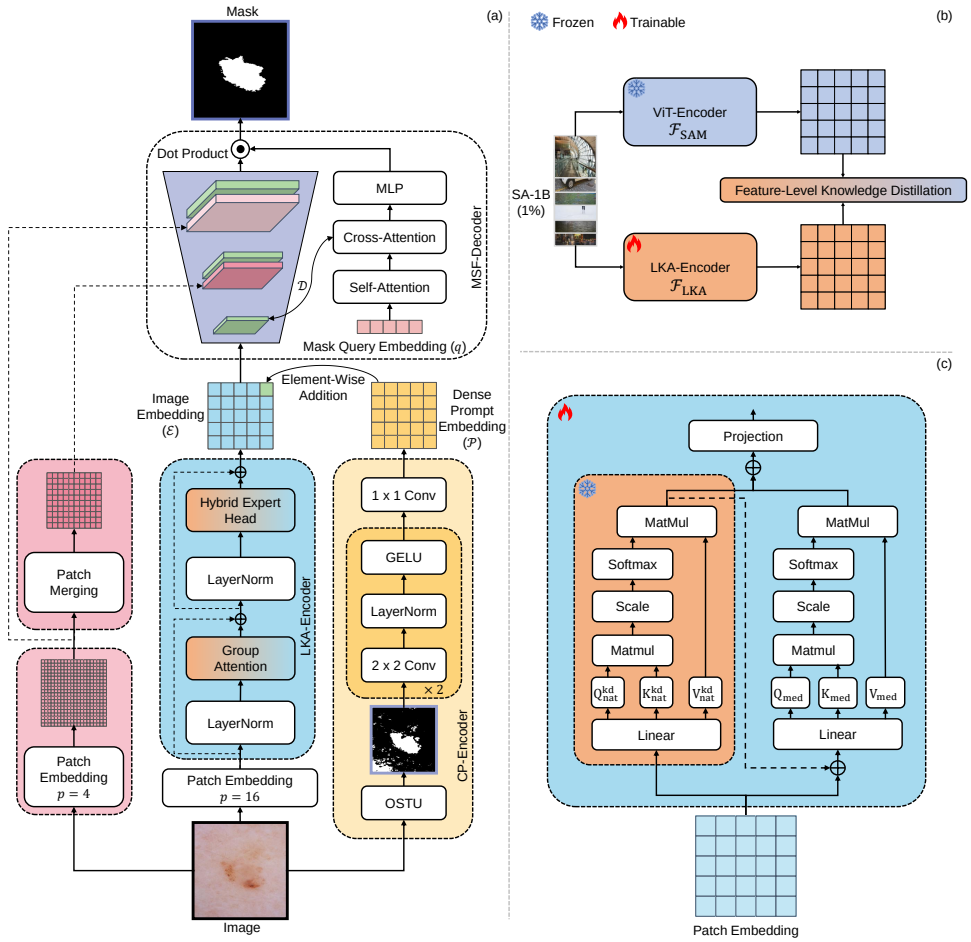

Figure 1: The overview of proposed Med-FastSAM framework. (a) Med-FastSAM contains three innovative components: LKA-Encoder, CP-Encoder and MSF-Decoder. (b) Feature-Level Distillation. (c) Group Attention.

U-Net++ [34] enhances feature propagation with denser skip connections, while Attention U-Net [23] incorporates attention mechanisms to improve focus on relevant regions. TransUNet [2] and Swin-UNet [1] introduce Transformer-based architectures to better handle long-range dependencies, and nnU-Net [11] dynamically adjusts its architecture to suit various tasks. ACC-UNet [10] further refines segmentation accuracy through additional attention mechanisms. Despite these improvements, many of these models have not been extensively trained on large-scale datasets, which limits their generalizability. Our Med-FastSAM model addresses these generalization issues by optimizing the architecture and training strategies, enabling superior performance across diverse medical imaging datasets.

2.2 Segment Anything in Medical Image Segmentation

The Segment Anything Model (SAM) [13] has demonstrated exceptional flexibility in handling a wide range of segmentation tasks, particularly in natural image processing. The robust ViT encoder of SAM, coupled with a prompt encoder and versatile decoder, has made it adaptable to various challenges, including medical image segmentation. Models like SAMMI [9] and MedSAM [20] apply global fine-tuning of SAM to multiple medical datasets, achieving strong performance but at a high computational cost. To reduce this burden, models such as SAMed [33] and Med-SA [28] employ techniques like Adapters and LoRA for parameter-efficient fine-tuning, reducing the number of trainable parameters. However, they still maintain large overall model sizes due to the extensive architecture of the ViT encoder. Furthermore, reliance on manual prompts in these models introduces

challenges in clinical settings, where manual annotations are time-consuming and can introduce bias. In contrast, our Med-FastSAM model addresses these limitations by achieving automatic prompt generation, eliminating the need for manual annotations while maintaining a smaller and more efficient model architecture, making it more practical for clinical use.

3 Method

The architecture of our Med-FastSAM is illustrated in Fig 1a. It is mainly composed of LKA-Encoder, CP-Encoder and MSF-Decoder. Given a medical image, it is first fed into the LKA-Encoder, extracting hybrid image embeddings. Meanwhile, the CP-Encoder processes the image to generate a coarse prompt for assisting segmentation decoding. Then, the MSF-Decoder incorporates the image embedding and local fine-grained features at different scales to predict segmentation masks.

3.1 Lightweight Knowledge Aggregation Encoder

The standard SAM contains a large image encoder that costs huge computational resources, degrading its applicability in real-world scenarios [31]. Although existing parameter-efficient fine-tuning techniques reduce parameters in the training phase, they still rely on the large ViT [5] encoder and additional trainable parameters increase the size of the entire model [7, 8]. To address this issue, we propose the Lightweight Knowledge Aggregation Encoder (LKA-Encoder) that utilizes knowledge distillation to reduce memory costs of SAM-based knowledge and leverages a set of learnable parameters for efficient fine-tuning in medical scenarios. Specifically, the image encoder of SAM [13] $\mathcal{F}_{\mathrm{SAM}}$ mainly involves a set of pre-trained functions $\{Q_{\mathrm{nat}}(\cdot), K_{\mathrm{nat}}(\cdot), V_{\mathrm{nat}}(\cdot), \Psi_{\mathrm{nat}}(\cdot)\} \subset \mathcal{F}_{\mathrm{sam}}$, where $Q_{\mathrm{nat}}(\cdot)$, $K_{\mathrm{nat}}(\cdot)$ and $V_{\mathrm{nat}}(\cdot)$ stand for the *query*, *key* and *value* branches of the multi-head attention layer and $\Psi_{\mathrm{nat}}(\cdot)$ is a MLP layer. To reduce computational costs and inspired by [], we adopt a feature-level knowledge distillation strategy to project the weight of these functions to a lower-dimensional space d. In progress, the loss function $\mathcal{L}_{\mathrm{FLKD}}$ is defined as:

$$\mathcal{L}_{\mathrm{FLKD}} = \frac{1}{N}\sum_{z=1}^{N} \left|\left|\mathcal{F}_{\mathrm{SAM}}(u_z) - \mathcal{F}_{\mathrm{LKA}}(u_z)\right|\right|_2^2, \tag{1}$$

where u_z is a natural image sampled from 1% SA-1B dataset and $\mathcal{F}_{\mathrm{LKA}}(\cdot)$ represents our LKA-Encoder. Let the input patch embeddings be denoted by $\mathcal{X} \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times d}$, where H, W, p and d stand for height, width, patch size and number of channels, respectively. In LKA-Encoder, we devise a Group Attention (GA) inspired by the split attention of CNN. As presented in Fig. 1c, the first branch performs a general self-attention computation:

$$\mathcal{A}_{\mathrm{nat}} = \mathrm{softmax}(\frac{Q^{\mathrm{kd}}_{\mathrm{nat}}(\mathcal{X}) \cdot K^{\mathrm{kd}}_{\mathrm{nat}}(\mathcal{X})^T}{\sqrt{d}}) \cdot V^{\mathrm{kd}}_{\mathrm{nat}}(\mathcal{X}), \tag{2}$$

where $\mathcal{A}_{\mathrm{nat}} \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times d}$, $\{Q^{\mathrm{kd}}_{\mathrm{nat}}(\cdot), K^{\mathrm{kd}}_{\mathrm{nat}}(\cdot), V^{\mathrm{kd}}_{\mathrm{nat}}(\cdot)\} \subset \mathcal{F}_{\mathrm{LKA}}$ includes the knowledge of natural images distilled from Eq. 2. The second branch contains a set of learnable parameters $\{Q_{\mathrm{med}}(\cdot), K_{\mathrm{med}}(\cdot), V_{\mathrm{med}}(\cdot)\}$ to adapt the attention map from natural to medical domains in a cascaded manner:

$$\mathcal{A}_{\mathrm{med}} = \mathrm{softmax}(\frac{Q_{\mathrm{med}}(\mathcal{X} \oplus \mathcal{A}_{\mathrm{nat}}) \cdot K_{\mathrm{med}}(\mathcal{X} \oplus \mathcal{A}_{\mathrm{nat}})^T}{\sqrt{d}}) \cdot V_{\mathrm{med}}(\mathcal{X} \oplus \mathcal{A}_{\mathrm{nat}}), \tag{3}$$

where $\oplus$ stands for the operation of element-wise addition. Then, $\mathcal{A}_{\mathrm{med}}$ integrates with $\mathcal{A}_{\mathrm{nat}}$ to generate rich image embeddings $\mathcal{E}$ that precisely focuses on the disease region:

$$\mathcal{E} = \mathcal{F}_{\mathrm{linear}}(\mathcal{A}_{\mathrm{nat}} \oplus \mathcal{A}_{\mathrm{med}}), \tag{4}$$

where $\mathcal{F}_{\mathrm{linear}}(\cdot)$ is a projection layer. Furthermore, we devise a Hybrid Expert Head (HEH) to extend the model capacity without unduly increasing computational overheads. It contains two parallel MLP layers $\{\Psi^{\mathrm{kd}}_{\mathrm{nat}}(\cdot), \Psi_{\mathrm{med}}(\cdot)\}$. Similarly to GA, we compress the common knowledge from $\Psi_{\mathrm{nat}}(\cdot)$ to $\Psi^{\mathrm{kd}}_{\mathrm{nat}}(\cdot)$. $\Psi_{\mathrm{med}}(\cdot)$ aims to learn medical domain-specific knowledge. Then, we aggregate both outputs with the residual style to update the image embeddings, as follows:

$$\mathcal{E} \leftarrow \mathcal{E} + \Psi^{\mathrm{kd}}_{\mathrm{nat}}(\mathcal{E}) + \Psi_{\mathrm{med}}(\mathcal{E}). \tag{5}$$

On this basis, the proposed LKA-Encoder overcomes the demand for fine-tuning the large vision transformer and achieves efficient transfer learning between natural and medical images.

3.2 Coarse Prompt Encoder

The standard SAM mainly requires manual annotations as prompts[3] to assist segmentation decoding, which is time-consuming and expensive, limiting its applicability for clinical scenarios. To overcome this challenge, we introduce the Coarse Prompt Encoder (CP-Encoder) that leverages Ostu thresholding method to automatically generate coarse segmentation masks as prompts. Specifically, Otsu is a traditional image processing method that automatically determines the optimal threshold to separate the foreground from the background. The algorithm starts by computing the histogram $h(\cdot)$ of the grayscale image I:

$$h(i) = \frac{1}{\mathcal{N}} \sum_{x=1}^{H} \sum_{y=1}^{W} \delta(I(x,y) - i), \tag{6}$$

where δ is the Dirac delta function, $\mathcal{N}$ is the total number of pixels in the image, x and y are the pixel coordinates. Then, the between-class variance σ_B^2 for each threshold t is defined as:

$$\sigma_B^2(t) = \frac{(\sum_{i=0}^{L-1} i \cdot h(i) \cdot \sum_{i=0}^{t} h(i) - \sum_{i=0}^{t} i \cdot h(i))^2}{\sum_{i=0}^{t} h(i) \cdot (1 - \sum_{i=0}^{t} h(i)) + \epsilon}, \tag{7}$$

where L represents the number of gray levels in the image, ϵ is a small constant to avoid division by 0. The optimal threshold t^* is found by:

$$t^* = \arg\max_t \sigma_B^2(t). \tag{8}$$

The threshold value is then scaled back to the range $[0, 1]$. We discretise the values of the image to either 0 or 1 using this threshold t^*. The coarse mask $\mathcal{C}$ is generated as follows:

$$\mathcal{C}(x,y) = \begin{cases} 1, & \text{if } I(x,y) \geq t^* \\ 0, & \text{if } I(x,y) < t^* \end{cases} \tag{9}$$

On this basis, we receive a coarse mask $\mathcal{C}$ corresponding to the input image. To convert this coarse mask into a set of dense prompts $\mathcal{P}$, we use a two-layer convolutional neural network (CNN) to perform the downsampling operation:

$$\mathcal{P} = \mathcal{F}_{\text{conv}}^{1\times1}(\sigma(\mathcal{F}_{\text{LN}}(\mathcal{F}_{\text{conv}}^{2\times2}(\sigma(\mathcal{F}_{\text{LN}}(\mathcal{F}_{\text{conv}}^{2\times2}(\mathcal{C})))))), \tag{10}$$

where $\mathcal{F}_{\text{conv}}^{2\times2}(\cdot)$ is a 2×2 convolution with stride 2, $\mathcal{F}_{\text{LN}}(\cdot)$ is LayerNorm, σ stands for the GELU activation function and $\mathcal{F}_{\text{conv}}^{1\times1}(\cdot)$ is a 1×1 for aligning the channel with the image embeddings $\mathcal{E}$. Overall, the proposed GKP-Encoder can produce a set of sufficient semantic prompt tokens to promote medical image segmentation and eliminate the need for our Med-FastSAM on manual annotations.

3.3 Multi-Scale Feature Decoder

The mask decoder of SAM [13] directly utilises, the tokens from large-size, patches to predict masks, resulting in the loss of fine-grained details crucial for accurate medical image segmentation. To address this issue, we propose the Multi-Scale Feature Decoder (MSF-Decoder) that provides additional semantic information at different scales. Specifically, we first adopt the self-attention mechanism to update the mask query embedding q. Then, we combine the image embedding with the dense prompt embedding and conduct cross-attention with q:

$$\mathcal{D} = \text{softmax}(\frac{((\mathcal{E} \oplus \mathcal{P}) + \phi) \cdot q^T}{\sqrt{d}}) \cdot q \oplus (\mathcal{E} \oplus \mathcal{P}), \tag{11}$$

where $\mathcal{D}$ is the updated embedding. ϕ represents the positional encoding and $\cdot$ is the matrix multiplication. Inspired by the U-shape architecture [24], for the input image, we additionally perform patch embedding on the input images with different patch sizes p and then conduct multi-scale patch merging. The generated two multi-scale embeddings are denoted as r_1 and r_2. We combine these low-level semantic information maps with the updated embedding $\mathcal{D}$, constructing a hierarchical decoding workflow to predict segmentation mask $\mathcal{M}$ as follows:

$$\mathcal{M} = \sigma(\mathcal{F}_{\text{up}}^{2\times2}\sigma((\mathcal{F}_{\text{LN}}(\mathcal{F}_{\text{up}}^{2\times2}(\mathcal{D}) \frown r_1) \frown r_2))), \tag{12}$$

$$\mathcal{M} \leftarrow \mathcal{F}_{\text{inter}}(\Phi(\mathcal{M} \cdot \Psi_{\text{query}}(q))), \tag{13}$$

where $\mathcal{F}_{\text{up}}^{2\times2}(\cdot)$ stands for the upsampling operation with 2×2 kernel size (e.g. transpose convolution), $\frown$ is the concatenation operation, Φ is the sigmoid operation and $\mathcal{F}_{\text{inter}}(\cdot)$ is a bilinear interpolation function to recover the shape of segmentation masks. The predicted segmentation mask is supervised by the weighted combination of focal loss [19] $\mathcal{L}_{\text{focal}}$ and dice loss $\mathcal{L}_{\text{dice}}$, as follows:

$$\mathcal{L}_{\mathcal{M}} = \lambda\mathcal{L}_{\text{focal}} + (1-\lambda)\mathcal{L}_{\text{dice}}, \tag{14}$$

where λ is the coefficient to balance the weight of these two loss terms. Overall, our MSF-Decoder utilises convolutions to capture local fine-grained details, improving the precision of the segmentation mask.

4 Experiment

4.1 Datasets

To evaluate the effectiveness of the proposed Med-FastSAM framework, we first train our model on the ISIC-2018 [4, 26] and MoNuSeg-2018 [14] datasets and follow their official guidelines to construct training, validation and test sets. We further select two external datasets: PH2 [21] and TNBC [22] as the unseen target domains. All images from the PH2 [21] and TNBC [22] datasets are used for testing domain generalisation capabilities of models. The details are as follows:

4.1.1 ISIC-2018

The ISIC-2018 [4, 26] dataset focuses on dermoscopy images for skin lesion segmentation, which is crucial for melanoma detection. It includes 2594 training images, 100 validation images, and 1000 test images with different image sizes. The dataset covers various body portions such as the back, arms, and legs.

4.1.2 MoNuSeg-2018 Dataset

The MoNuSeg-2018 [14] dataset consists of histopathology images for nuclei segmentation. It is collected from the liver, breast, colon, stomach, bladder, kidney and prostate organs. The dataset contains 30 training images and 14 test images with the fixed image size of 1000×1000.

4.1.3 PH2 Dataset

The PH2 [21] dataset includes 200 dermoscopy images of various pigmented skin lesions, such as nevi and melanomas, primarily from the back and limbs. It serves as a benchmark for evaluating lesion segmentation and diagnosis. All images have the same resolution of 767×576.

4.1.4 TNBC Dataset

The TNBC [22] dataset contains 50 histopathology images of 512×512 sampled from triple-negative breast cancer tissues. This dataset is particularly challenging due to the dense clustering and variability of nuclei, making it essential for evaluating models in complex cancer pathology.

4.2 Implementation Details

All experiments are conducted with PyTorch 1.13.0 framework on a single NVIDIA RTX8000 (48GB) Tensor Core GPU, 64-core CPU, and 520G RAM. We set batch sizes and epochs to 4 and 200 respectively. An Adam optimizer with a learning rate of 5e-4 is used for training and the loss coefficient λ is set as 0.8. We resize the original images into 1024×1024. Inspired by [30, 29], we set d as 320 and p as 16 to construct our Med-FastSAM. For a fair comparison, all baseline models use the same training configuration as our framework and all SAM models use ViT-B [5] structure as the image encoder. To compare with the point prompt mode [13], these models apply the *ConnectedComponentsWithStats* function in OpenCV to calculate the centroid of each object (e.g. lesion, nuclei) as point prompts [9].

Table 1: Comparison with state-of-the-arts on two *source* domains.

Methods	Manual Prompt	Tuned/ Total (M)	ISIC-2018		MoNuSeg-2018	
			mIoU(%)	Dice(%)	mIoU(%)	Dice(%)
U-Net [24]	✘	13.40/13.40	74.66	83.26	60.12	74.53
ACC-UNet [10]		16.68/16.68	75.89	84.72	64.06	77.76
nnU-Net [11]		30.60/30.60	78.21	86.43	67.52	80.52
SAM [13]	Point	4.06/93.74	76.41	85.34	66.66	79.93
SAMMI [9]		4.06/93.74	78.35	86.47	67.23	80.32
MedSAM [20]		4.06/93.74	76.93	85.61	62.23	76.32
Med-SA [28]		7.10/100.84	79.02	86.97	68.53	81.26
SAMed [33]		4.21/93.88	78.98	86.93	68.00	80.88
MobileSAM [32]		10.13/10.13	78.46	86.65	66.28	79.64
EfficientSAM [30]		25.38/25.38	76.87	85.53	67.80	80.72
RepViT-SAM [27]		9.98/9.98	77.01	85.72	64.40	78.29
Med-FastSAM	✘	8.62/14.48	**80.38**	**87.84**	**69.35**	**81.75**

Table 2: Comparison with state-of-the-arts on two *target* domains.

Methods	Manual Prompt	ISIC-2018 ⇒ PH2		MoNuSeg-2018 ⇒ TNBC	
		mIoU(%)	Dice(%)	mIoU(%)	Dice(%)
U-Net [24]	✘	77.56	86.70	36.19	50.02
ACC-UNet [10]		78.12	86.99	39.34	52.43
nnU-Net [11]		79.51	88.03	41.72	53.81
SAM [13]	Point	81.65	89.16	45.34	59.76
SAMMI [9]		81.93	89.29	46.16	60.84
MedSAM [20]		81.27	88.94	43.55	56.98
Med-SA [28]		82.71	90.32	46.65	60.92
SAMed [33]		82.54	90.10	46.28	60.73
MobileSAM [32]		81.36	89.01	45.17	59.53
EfficientSAM [30]		80.29	88.59	45.91	60.25
RepViT-SAM [27]		81.15	88.87	44.93	59.31
Med-FastSAM	✘	**83.48**	**90.62**	**47.89**	**62.26**

Table 3: Ablation study of Med-FastSAM on the MoNuSeg dataset.

LKA-Encoder		CP-Encoder	MSF-Decoder	mIoU(%)	Dice(%)	Param(M)	FPS
HEH	GA						
				48.36	64.92	93.74	4.53
✓				52.06	67.94	11.32	26.68
✓	✓			68.36	80.82	14.46	24.19
✓	✓	✓		68.68	81.26	14.47	24.09
✓	✓	✓	✓	69.53	81.53	14.48	21.18

4.3 Comparison with State-of-the-Arts on Source Domains

To evaluate our Med-FastSAM in medical image segmentation, we conduct the comparison with state-of-the-arts in skin lesion and nuclei segmentation tasks. Specifically, we select U-shape architectures (i.e., U-Net [24], ACC-UNet [10] and nnU-Net [11]), SAM [13], medical SAMs (i.e., SAMMI [9] and MedSAM [20]), PEFT SAMs (i.e., Med-SA [28] and SAMed [33]) and lightweight SAMs (i.e., MobileSAM [32], EfficientSAM [30] and RepViT-SAM [27]) as baselines. We first train all models on ISIC-2018 [4, 26] and MoNuSeg-2018 [14] datasets, respectively. As shown in Table 1, among the U-Net variants, nnU-Net achieves the best performance on ISIC-2018 with the Dice of 86.43% and

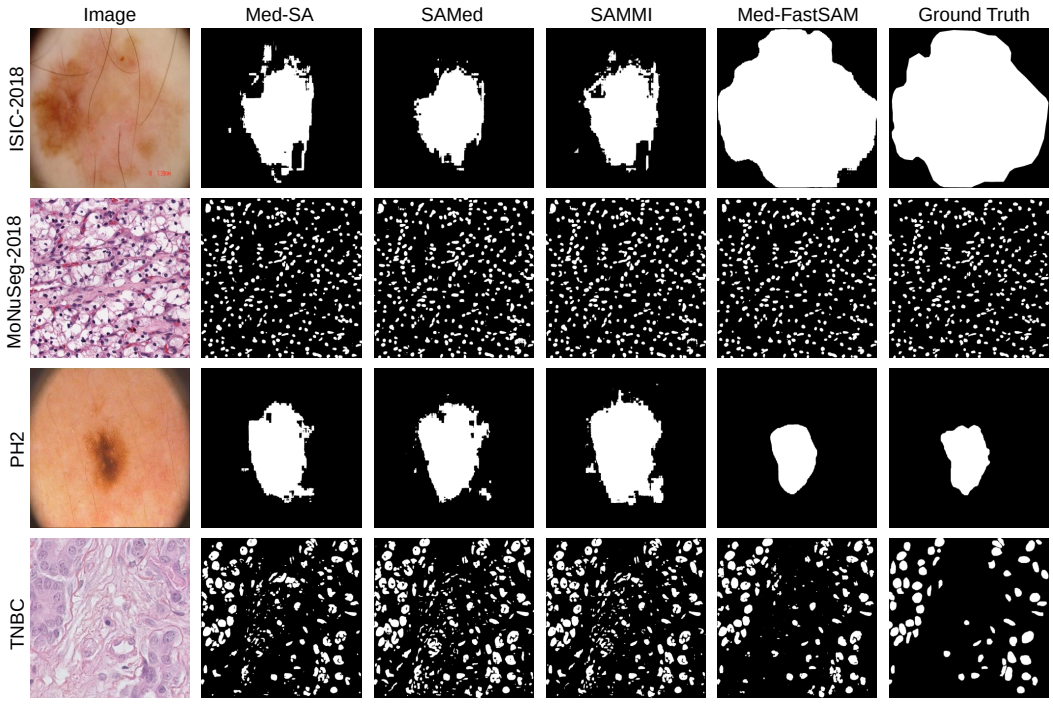

Figure 2: Qualitative comparison on two source domains: ISIC-2018 and MoNuSeg-2018 and two target domains: PH2 and TNBC.

80.52% while the proposed Med-FastSAM surpasses it with an improvement of 1.41% and 1.23%. For foundation models, PEFT SAMs show better performance than Medical SAMs and lightweight SAMs. In contrast, our Med-FastSAM framework outperforms these state-of-the-arts with the Dice of 87.84% and 81.75%, which is 0.87% and 0.49% higher than Med-SA. Notably, Med-FastSAM does not require any manual annotations as prompts, which is more clinical-friendly.

Furthermore, we report the tuned and total parameters for each segmentation architecture. It is observed that the total parameters of U-Net variants are lower than SAM-based foundation models, e.g., nnU-Net uses 32.64% parameters compared to SAM. On the other hand, foundation models have the ability to provide generalised feature representations as they are pre-trained on a large-scale dataset. Therefore, existing methods adopt the decoder-only strategy or PEFT techniques to adapt SAM from natural to medical domains while reducing training complexity, e.g., Med-SA only tunes 7.04% of all parameters but the size of the final model does not decrease, which is still expensive. Lightweight SAMs transfer the knowledge of SAM to small models, compressing the model size but their performance is inferior to PEFT SAMs. On the contrary, our Med-FastSAM frameworks use only 15.45% parameters compared to the standard SAM and outperform state-of-the-arts on two datasets, achieving significant generalisation-efficiency trade-offs.

4.4 Comparison with State-of-the-Arts on Unseen Target Domains

In this section, we demonstrate the domain generalisation capabilities of Med-FastSAM on two external datasets. Specifically, all trained models on ISIC-2018 and MoNuSeg-2018 datasets are respectively evaluated on PH2 and TNBC datasets without any further fine-tuning. The results are presented in Table 2. We can observe that all SAM-based frameworks perform better than U-Net variants due to their larger image encoder and extra manual prompts, e.g., SAM achieves the Dice of 89.16% and 59.76%, which is 1.13% and 5.95% higher than nnU-Net. Lightweight SAMs are inferior to PEFT SAMs as the full-parameter fine-tuning method loses the common knowledge pre-trained on the large-scale dataset. In contrast, our Med-FastSAM adopts a semi-parameter fine-tuning method to take advantage of foundation models and automatic prompt generation to achieve superior domain generalisation capabilities with the Dice increase of 0.3% and 1.34% on two datasets compared to Med-SA. To perform the qualitative comparison with state-of-the-arts, we

provide the visualisation results in Figure 2. It can be revealed that our Med-FastSAM generates the best segmentation results with fewer false positives, especially for skin lesion segmentation. Overall, the proposed Med-FastSAM illustrates remarkable domain-generalised medical image segmentation without laborious annotations as prompts.

4.5 Ablation Study

In this section, a detailed ablation study is conducted on the MoNuSeg-2018 dataset to evaluate the efficiency of three components, including the LKA-Encoder, CP-Encoder and MSF-Decoder, in Med-FastSAM, which is provided in Table 3. We consider the original SAM [13] as the baseline. Firstly, introducing the LKA-Encoder significantly increases the mIoU from 48.36% to 68.36% and the Dice score from 64.92% to 80.32% by providing enhanced feature representation. Meanwhile, it reduces the number of parameters from 93.74M to 14.46M and increases the inference speed of Med-FastSAM from 4.53 FPS to 24.19 FPS. Secondly, the CP-Encoder eliminates the requirement of manual prompts and further boosts model performance by effectively incorporating coarse prompt information. Finally, the MSF-Decoder leverages multi-scale features to further improve the prediction accuracy of segmentation masks, with the mIoU of 69.53% and the Dice score of 81.53%. This comprehensive evaluation demonstrates the effectiveness of each module in the proposed Med-FastSAM.

5 Conclusion

In this paper, we propose Med-FastSAM to enhance the transfer efficiency of SAM for domain-generalised medical image segmentation. Med-FastSAM integrates three key modules: LKA-Encoder improves feature representation and reduces computational costs through feature-level knowledge distillation and semi-parameter fine-tuning strategies; CP-Encoder eliminates the reliance on manual annotations by incorporating coarse prompt information, enabling fully automated segmentation; and the MSF-Decoder captures local fine-grained details by leveraging multi-scale features. Extensive experiments confirm the superiority of Med-FastSAM over existing medical SAM models, demonstrating its enhanced generalisation capability on unseen domains. Future research will optimise Med-FastSAM to accommodate diverse medical imaging modalities.

References

[1] 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 Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham, 2023. Springer Nature Switzerland.

[2] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.

[3] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[4] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, 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] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[6] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023.

[7] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799. PMLR, 2019.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[9] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Med. Image Anal.*, 92:103061, 2024.

[10] Nabil Ibtehaz and Daisuke Kihara. Acc-unet: A completely convolutional unet model for the 2020s. In *MICCAI*, pages 692–702. Springer, 2023.

[11] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[12] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *NeurIPS*, 36, 2024.

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, October 2023.

[14] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Trans. Med. Imaging*, 36(7):1550–1560, 2017.

[15] Chenxin Li, Xin Lin, Yijin Mao, Wei Lin, Qi Qi, Xinghao Ding, Yue Huang, Dong Liang, and Yizhou Yu. Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in biology and medicine*, 141:105144, 2022.

[16] Chenxin Li, Xinyu Liu, Cheng Wang, Yifan Liu, Weihao Yu, Jing Shao, and Yixuan Yuan. Gtp-4o: Modality-prompted heterogeneous graph learning for omni-modal biomedical representation. *arXiv preprint arXiv:2407.05540*, 2024.

[17] Chenxin Li, Wenao Ma, Liyan Sun, Xinghao Ding, Yue Huang, Guisheng Wang, and Yizhou Yu. Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. *Neural Computing and Applications*, pages 1–14, 2022.

[18] Chenxin Li, Yunlong Zhang, Zhehan Liang, Wenao Ma, Yue Huang, and Xinghao Ding. Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 61–65. IEEE, 2021.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[20] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[21] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *EMBC*, pages 5437–5440. IEEE, 2013.

[22] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Trans. Med. Imaging*, 38(2):448–459, 2018.

[23] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, M. J. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[25] Liyan Sun, Chenxin Li, Xinghao Ding, Yue Huang, Zhong Chen, Guisheng Wang, Yizhou Yu, and John Paisley. Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Computers in biology and medicine*, 140:105067, 2022.

[26] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[27] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *CVPR*, pages 15909–15920, 2024.

[28] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[29] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *ECCV*, pages 68–85. Springer, 2022.

[30] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *CVPR*, 2024.

[31] Qing Xu, Wenwei Kuang, Zeyu Zhang, Xueyao Bao, Haoran Chen, and Wenting Duan. Sppnet: A single-point prompt network for nuclei image segmentation. In *MLMI*, pages 227–236. Springer, 2023.

[32] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[33] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[34] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer International Publishing.