Automating MedSAM by Learning Prompts with Weak Few-Shot Supervision

Mélanie Gaillochet[1,2,3], Christian Desrosiers[1], and Hervé Lombaert[1,2,3]

[1] ÉTS Montréal, Canada
[2] Polytechnique Montréal, Canada
[3] Mila - Quebec AI Institute, Université de Montréal, Canada

Abstract. Foundation models such as the recently introduced Segment Anything Model (SAM) have achieved remarkable results in image segmentation tasks. However, these models typically require user interaction through handcrafted prompts such as bounding boxes, which limits their deployment to downstream tasks. Adapting these models to a specific task with fully labeled data also demands expensive prior user interaction to obtain ground-truth annotations. This work proposes to replace conditioning on input prompts with a lightweight module that directly learns a prompt embedding from the image embedding, both of which are subsequently used by the foundation model to output a segmentation mask. Our foundation models with learnable prompts can automatically segment any specific region by 1) modifying the input through a prompt embedding predicted by a simple module, and 2) using weak labels (tight bounding boxes) and few-shot supervision (10 samples). Our approach is validated on MedSAM, a version of SAM fine-tuned for medical images, with results on three medical datasets in MR and ultrasound imaging. Our code is available on https://github.com/Minimel/MedSAMWeakFewShotPromptAutomation.

Keywords: Large Vision Models · Segmentation · Medical · Prompt.

1 Introduction

Annotation is a well-known labour-intensive and time-consuming task in medical imaging. Supervised segmentation models trained to identify specific regions of interest do not generalize well to new domains or classes and require more data and retraining when considering a new task. This increases the cost of developing segmentation models to solve multiple tasks. The need for universal models that can be applied to various tasks after training has hence been growing in medical image analysis. The introduction of foundation models for image segmentation such as the recent Segment Anything Model (SAM) [10], as well as its versions adapted for medical imaging [21], notably MedSAM [12], have appeared as a game-changer in the field of computer vision and medical image analysis. These models have shown remarkable performance on a variety of segmentation tasks. However, they remain promptable models that require user interaction to

obtain the segmentation mask of a target object. Furthermore, their zero-shot performance depends on the quality of the user prompt. This reliance on user interaction hinders their integration into automatic pipelines and limits their usability at a large scale.

Recent attempts have been made to automate the prompt generation of SAM [20,17,22]. However, these methods typically require samples with ground-truth segmentation masks, which are costly to obtain in the medical domain.

This paper proposes a lightweight add-on prompt module which learns to generate prompt embeddings directly from SAM's image embedding. Our end-to-end approach enables SAM models to specialize on the segmentation of a specific region and only requires few weakly-annotated samples. This reduces the interaction cost of developing specialized segmentation models. Our validation shows that, given only few training samples weakly annotated with tight boxes, promptable foundation model can effectively generate segmentation masks of target regions without requiring manual prompt inputs.

Foundation models for medical image segmentation. Vision foundation models have achieved tremendous success in computer vision tasks thanks to large-scale pre-training. In particular, the Segment Anything Model (SAM) [10], based on vision transformers [5] and trained on 1B masks and 11M images, was recently introduced as a prompt-driven foundation model for segmentation. Trained on natural images, SAM obtains uneven performances on medical data [7,13,18], inducing its adaptation to the medical domain [4,12,17]. In particular, MedSAM [12], a foundation model for universal medical image segmentation was trained on 1.5 million image-mask pairs over 10 imaging modalities. These models provide impressive zero-shot performance, but remain promptable models that require user interaction at inference.

Prompt automation for SAM. Motivated by its performance in Natural Language Processing [2], prompt-tuning has successfully been applied to large vision models [8]. Hence, methods that have focused on specializing SAM, a promptable model, have naturally explored prompt generation. Given few fully labeled samples, the self-prompting unit of [20] automatically generates a real point and bounding box from SAM's image embedding. AutoSAM replaces SAM's prompt encoder with a Harmonic Dense Net to adapt segmentation to medical images [17]. A recent training-free approach, PerSAM [22], encodes positive-negative location priors as prompt tokens to produces automatic segmentations of a specific object from a single reference image and mask. As opposed to our approach, all of these methods require samples with full segmentation masks.

Segmentation with bounding box annotations. Bounding boxes have emerged as an alternative to onerous annotation masks. Most methods use bounding boxes as an initial pseudo-label of the target region. A classic iterative graph-cut-based algorithm, GrabCut [16], separates the foreground from its background given a bounding box. DeepCut [14] extends GrabCut to neural networks using existing heuristics. More recently, the bounding box tightness prior was adapted to deep learning-based models by imposing a set of constraints on the predictions [9], and was combined with multiple instance learning and

smooth maximum approximation [19].

Our contribution. This work aims to efficiently automate MedSAM, a variant of SAM for the medical domain, to segment any target region through the use of few, weakly-labeled samples. Our approach introduces an innovative improvement by substituting the original prompt encoder, which requires user input, with an enhanced lightweight adaptable prompt-learning module that:

1. Automatically **generates a prompt embedding** from the input image
2. Trains with only **weak labels** (tight bounding boxes) and **few-shot** learning
3. Is easily added on top of MedSAM (no fine-tuning)

The next sections present our proposed prompt module for MedSAM and demonstrate its usefulness on various medical image segmentation tasks.

2 Methodology

2.1 Preliminaries: MedSAM architecture

Our approach builds upon on MedSAM [12], a variant of SAM [10] fine-tuned on medical data. The model has three main components: a large image encoder E_{img}, a prompt encoder E_{pr} and a lightweight mask decoder D_{mask}.

While the image encoder computes an embedding of the input image x, the prompt encoder outputs two sparse and dense embeddings from the provided set of prompts $[pr]$, respectively points or bounding boxes (BB), and a mask. The network produces a probability map f_θ by taking x and a prompt embedding $Z_{pr} = E_{pr}([pr])$:

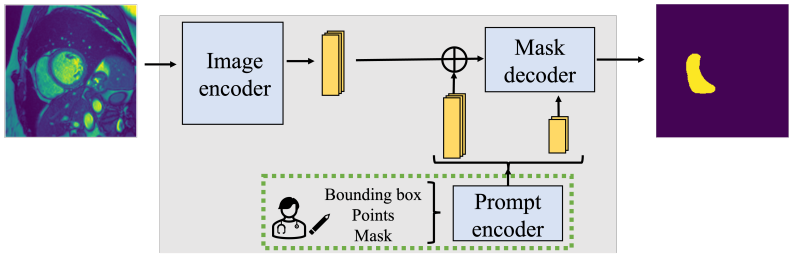
$$f_\theta = \sigma\big(D_{mask}(E_{img}(x),\, Z_{pr})\big),$$

where σ is the sigmoid function.

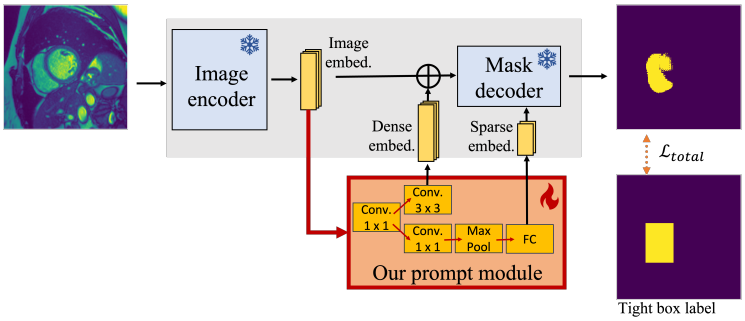
We present an end-to-end approach to remove the typical dependence on user-defined prompts $[pr]$, without modifying the pretrained MedSAM network.

2.2 Lightweight prompt module

Our approach consists of a prompt module trained to compute directly Z_{pr} from the image embedding provided by MedSAM (see Fig.1b). The module outputs two embeddings of the same shape as those generated by MedSAM (Fig.1a). Originally, the dense prompt embedding has a spatial correspondence with the image and can be considered as a low-quality segmentation map, while the sparse embeddings are spatial encodings of coordinates. Therefore, our prompt module generates a dense embedding through a convolutional layer and a sparse embedding through a fully connected (FC) layer.



(a) Promptable MedSAM [12]



(b) Automatic MedSAM (ours)

Fig. 1: Comparison between (a) MedSAM and (b) our automation of MedSAM via a learnt prompt module. Our prompt module replaces MedSAM's prompt encoder and learns to generate a relevant prompt embedding from the image embedding. Training employs losses that utilize only tight box labels.

2.3 Learning with tight box annotations

Denote as $X : \Omega \subset \mathbb{R}^{3\times H\times W} \rightarrow \mathbb{R}$ a 3-channel input image of height H and width W, where Ω is the spatial domain corresponding to each channel of the image. Moreover, let $Y \in \{0,1\}^{\Omega}$ be the ground-truth binary segmentation mask of X. Suppose we only have access to a tight bounding box $\tilde{Y}$ of the target. Ω_I and Ω_O define the regions respectively inside and outside the bounding box such that $\Omega_I + \Omega_O = \Omega$. This leads to a constrained optimization problem [9] from the bounding box annotations $\tilde{Y}$.

Emptiness of Ω_O. Since the region in Ω_I defined by the bounding box must contain the target object, Ω_O must contain only foreground. Hence, we can apply a Cross-entropy loss for all pixels $p \in \Omega_O$:

$$\mathcal{L}_{empty} = -\sum_{p\in\Omega_O} \log(1 - f_\theta(p)). \tag{1}$$

Tight box constraint in Ω_I. The tightness of the bounding box indicates that at least one foreground pixel must cross every horizontal and vertical line of weak label $\tilde{Y}$. As in [9], we soften this condition by considering segments of

width w instead of individual lines and ensure differentiability by considering output probabilities instead of the prediction mask. The condition formalizes as:

$$\sum_{p \in s_l} f_\theta(p) \geq w, \quad \forall s_l \in S_L, \tag{2}$$

where S_L is the set of all vertical and horizontal segments of width w that make the bounding box $\tilde{Y}$. We convert the inequality constraints of (2) to a loss using a penalty function ψ_t, and obtain:

$$\mathcal{L}_{tightbox} = \sum_{s_l \in S_L} \psi_t\Big(w - \sum_{p \in s_l} f_\theta(p)\Big). \tag{3}$$

The penalty function can be modeled as a simple scaled ReLU function, i.e. $\psi_t(x) = t \cdot \max(0, x)$. In this work, we instead resort to a pseudo log-barrier function, which provides a more stable optimization under multiple competing constraints. As $t \to \infty$, function $\psi_t(x)$ behaves as a hard barrier where $\psi_t(x) = \infty$ if $x > 0$, else $\psi_t(x) = 0$. In our method, we found that using a fixed value of $t = 5$ worked best.

Foreground size constraint. The bounding box $\tilde{Y}$ also sets a limit on the target size of the prediction mask. Again, we consider output probabilities rather than individual predictions to ensure differentiability. By applying priors on the fraction $\epsilon \in [0, 1]$ of pixels from Ω_I that belong to the background, we get:

$$\epsilon_1 |\Omega_I| \leq \sum_{p \in \Omega} f_\theta(p) \leq \epsilon_2 |\Omega_I|. \tag{4}$$

As before, we employ $\psi_t(x)$ to convert these inequality constraints into the following loss:

$$\mathcal{L}_{size} = \psi_t\Big(\epsilon_1 |\Omega_I| - \sum_{p \in \Omega} f_\theta(p))\Big) + \psi_t\Big(\sum_{p \in \Omega} f_\theta(p) - \epsilon_2 |\Omega_I|\Big). \tag{5}$$

Given (1), (3) and (5), and weights λ_1 and λ_2, the final loss becomes:

$$\mathcal{L}_{total} = \mathcal{L}_{empty} + \lambda_1 \mathcal{L}_{tightbox} + \lambda_2 \mathcal{L}_{size}. \tag{6}$$

3 Results

3.1 Datasets

Our experiments validate our method on three public datasets: the Head Circumference dataset[4] (HC18) [6], the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation[5] (CAMUS) [11] and the Automated Cardiac Diagnosis Challenge[6] (ACDC) [1]. For both cardiac datasets, the end diastole images are

[4] https://hc18.grand-challenge.org/
[5] https://www.creatis.insa-lyon.fr/Challenge/camus/
[6] https://humanheart-project.creatis.insa-lyon.fr/database/

used. For HC18, we filter out samples with ground-truth masks that could not be automatically generated by OpenCV from the circumference annotations, and split the ultrasound dataset into 507 training, 77 validation and 148 test images. For CAMUS, we focus on the left ventricle (LV) and left atrium (LA) segmentation and use 50 images for validation, 100 images for testing and the remaining 350 images for training. For ACDC, we focus on the right ventricle (RV) and LV segmentation and use 10 patients for validation (78 images), 50 for testing (470 images)and the remaining 90 patients (765 images) for training. Each sample has a minimum foreground size for all experiments.

Following [12], our preprocessing includes clipping the intensity values of each 2D image (HC18, CAMUS) or each 3D volume (ACDC) between the 0.5^{th} and 99.5^{th} percentiles and rescaling them to the range [0, 255]. We also partition each 3D volume of ACDC into 2D image which we resample to a fixed 1mm $\times$ 1mm resolution. We center crop and pad each sample to size 640×640 (HC), 512×512 (CAMUS) or 256×256 (ACDC). To meet MedSAM's requirements, we resize all images to a fixed $3 \times 1024 \times 1024$ size before inputting them in the model.

3.2 Implementation details

Model. Our backbone model is MedSAM based on ViT-B, the smallest version of SAM. The backbone remains frozen during training. We keep our prompt module lightweight by using few layers. A 1×1 convolution first reduces the number of channels. Then, the dense embedding is obtained through a 3×3 convolution, while the sparse embedding is obtained through a 1×1 convolution followed by max pooling and a fully connected layer. All convolutional layers are followed by ReLU activation. Our prompt module has thus 2.4M trainable parameters.

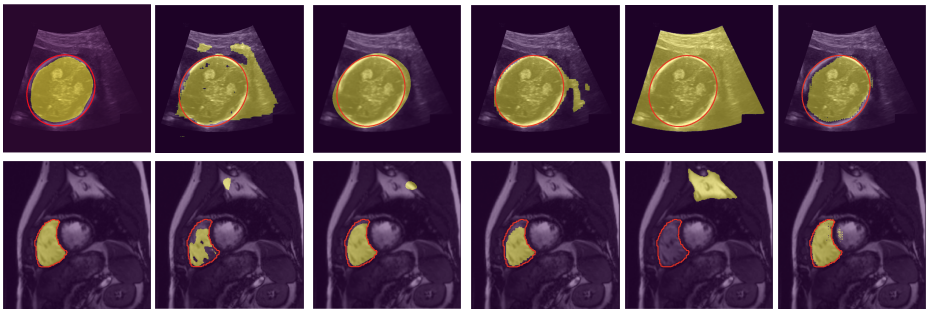
Loss parameters. We train our prompt module using $\mathcal{L}_{total}$, with $\lambda_1 = 0.0001$, $\lambda_2 = 0.01$. For our tight box constraint, we follow [9] and use segments of $w = 5$. We hypothesize that the foreground region is at least half the size its tight bounding box and set $[\epsilon_1, \epsilon_2] = [0.5, 0.9]$. Comparative training with full segmentation masks uses a Binary Cross-entropy Dice loss, each term having the same weight.

Training. We use a batch size of 4 and a learning rate (LR) of 0.001 with a multi-step scheduler decreasing LR by 0.1 after half the epochs and a weight decay of 0.0001. To minimize computational complexity, we do not use data augmentation. This allows us to discard MedSAM's image and prompt encoders after saving the image embeddings during an initial iteration, reducing the number of total parameters from 96.1M to only 6.5M (2.4M trainable). In the 10-shot setting, we repeat the experiments 9 times, with 3 initialization seeds and 3 training subsets selected uniformly at random. The results are averaged over these experiments. All experiments are implemented in Python 3.8.10 with Pytorch on NVIDIA RTX-A6000 GPUs.

Baselines. For each class, we compare our method with two specialized single-task models: a standard UNet [15] and a TransUNet [3]. We also validate our method against PerSAM [22] which automates SAM with 1-shot supervision,

Table 1: Model performance on test sets in terms of mean (std) 2D Dice similarity score (↑). Best results in few-shot settings are highlighted in bold.

Type	Method	# Samples	Mask labels	BB labels	HC	CAMUS		ACDC	
						LV	LA	RV	LV
Promptable	MedSAM [12] (w/ tight box)	–	–	–	95.19	94.50	89.23	93.78	95.45
Automatic (fully trained)	UNet [15]	All	✓		86.53 ±0.55	89.93 ±0.01	74.77 ±0.78	89.55 ±0.23	94.83 ±0.13
		10	✓		61.79 ±3.10	75.09 ±3.69	46.29 ±3.64	40.85 ±1.66	59.96 ±0.91
	TransUNet [3]	All	✓		96.32 ±0.19	92.92 ±0.29	85.04 ±0.15	90.79 ±0.07	94.08 ±0.07
		10	✓		**92.15** ±0.40	87.32 ±0.45	66.51 ±2.28	**68.69** ±0.58	78.98 ±1.36
Automatic (adapted)	AutoSAM [17]	All	✓		97.42 ±0.04	93.59 ±0.03	85.60 ±0.89	89.57 ±0.54	95.18 ±0.11
		10	✓		90.64 ±1.84	86.98 ±0.67	67.09 ±4.82	68.33 ±3.21	**84.17** ±2.05
	PerSAM [22]	1	✓		58.98 ±0.19	36.13 ±0.00	14.19 ±0.02	27.64 ±9.48	45.43 ±5.47
	Ours	All		✓	92.88 ±1.27	88.86 ±1.42	79.82 ±0.74	76.97 ±1.02	86.91 ±2.08
		10		✓	85.23 ±0.55	**88.38** ±0.83	**73.56** ±0.57	58.96 ±2.28	80.37 ±1.59



(a) MedSAM (Prompted) (b) UNet (10 masks) (c) TransUNet (10 masks) (d) AutoSAM (10 masks) (e) PerSAM (1 mask) (f) Ours (10 BB)

Fig. 2: Predicted segmentations on test samples of HC18 (row 1) and the right ventricle in ACDC (row 2). From left to right, (a) MedSAM prompted with a tight box, (b-d) UNet, TransUNet and AutoSAM, trained with ground-truth masks, (e) PerSAM using one reference image with its ground-truth mask, and (f) our method trained on tight bounding boxes. All automatic methods are given for the 10-shot setting, except PerSAM, a 1-shot approach. Ground-truth annotation is drawn in red, with predicted segmentation mask overlayed in yellow.

and AutoSAM [17] which uses a Harmonic Dense Net (41.6M parameters) to learn the prompt embedding. We train the UNet, TransUNet and AutoSAM on full segmentation masks with a standard Cross-entropy Dice loss. To improve the performance of the baseline models, longer training is used (200 epochs) with a larger batch size of 24 for TransUNet (following [3]). Similarly, the best results for PerSAM are obtained with the ViT-H backbone. Results for MedSAM are also included when prompted with the tightest bounding boxes (no noise).

3.3 Validation on multiple medical segmentation tasks

The Dice similarity score (DSC) is used as our evaluation criteria. We evaluate our approach on three datasets: CAMUS, an internal validation set of MedSAM, and HC18 and ACDC, two datasets never seen by MedSAM during training. This allows us to verify the ability of the prompt module to effectively learn which region to segment in both in-domain and out-of-domain data. Training is performed on both the entire training set and the difficult 10-shot regime.

Our results are given in Table 1 and Fig.2. First, with only tight bounding box (BB) annotations, our approach trained on all samples is able to outperform a UNet trained on ground-truth segmentation masks for 2 different tasks (HC and LA). The most significant results are observed in the 10-shot setting. With only 1.3% (ACDC), 2% (HC18) and 2.9% (CAMUS) of the total training samples, our approach sees only a slight decrease in performance (except for RV segmentation) compared to the full-data setting. This contrasts with the considerable performance drop observed with UNet and TransUNet in multiple segmentation tasks, even when both methods are trained with ground-truth mask labels. Therefore, our prompt module-based approach is not only more computationally efficient to train than specialized models, but it also requires only weak labels and appears more robust in the few-shot setting. AutoSAM displays slightly better test dice scores than our approach, but AutoSAM requires full ground-truth masks and uses a much heavier model to learn the prompt, yielding a 3-fold increase in the training time. Additionally, PerSAM, which uses only one reference image, fails to generate convincing segmentations. Its poor performance on medical datasets may be due to the fact that PerSAM generates point prompts used by SAM, which are more likely to introduce ambiguity [12].

The benefits of our proposed methods are visually supported by Fig.2. Our module trained on 10 training samples and tight bounding boxes yields segmentations much more convincing than those produced by a UNet trained on ground-truth masks. Given little training data, the UNet hallucinates large regions in the background. Our approach is also able to generate segmentation masks more faithful to the ground-truth than AutoSAM and PerSAM, two existing prompt-based adaptation methods for SAM.

4 Conclusion

This work proposes to automate a prompt-based universal model, such as MedSAM, by generating task-specific prompt embeddings directly from the image embedding of the foundation model. Our add-on module that can be integrated directly into MedSAM removes its dependence on user inputs. More importantly, by applying tightness and size constraints, our module can be trained effectively with only bounding box annotations while keeping MedSAM frozen. Furthermore, our 10-shot experiments has demonstrated that the number of samples required to train the model could be considerably reduced without a substantial degradation of the model performance. By adding a lightweight prompt module that can be trained with only few weak labels, MedSAM can efficiently be automated for specific tasks with minimal annotation hurdles.

Disclosure of Interests. This work is supported by the Canada Research Chair on Shape Analysis in Medical Imaging, the Research Council of Canada (NSERC) and the Fonds de Recherche du Québec – Nature et Technologies (FRQNT).

References

1. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M.M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M.: Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 1877–1901 (2020)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation (2021), `http://arxiv.org/abs/2102.04306`
4. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: SAM-Med2D (2023), `http://arxiv.org/abs/2308.16184`
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR) (2021)
6. Heuvel, T.L.A.v.d., Bruijn, D.d., Korte, C.L.d., Ginneken, B.v.: Automated measurement of fetal head circumference using 2D ultrasound images. Plos One **13**(8), e0200412 (2018)
7. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Liu, S., Chi, H., Hu, X., Yue, K., Li, L., Grau, V., Fan, D.P., Dong, F., Ni, D.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
8. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual Prompt Tuning. In: European Conference on Computer Vision (ECCV). vol. 13693, pp. 709–727 (2022)
9. Kervadec, H., Dolz, J., Wang, S., Granger, E., Ayed, I.B.: Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In: Conference on Medical Imaging with Deep Learning (MIDL). pp. 365–381 (2020)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: International Conference on Computer Vision (ICCV). pp. 3992–4003 (2023)
11. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L., Bernard, O.: Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)

12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
13. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis **89**, 102918 (2023)
14. Rajchl, M., Lee, M.C.H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., Rueckert, D.: DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks. IEEE Transactions on Medical Imaging **36**(2), 674–683 (2017)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241 (2015)
16. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23**(3), 309–314 (2004)
17. Shaharabany, T.: AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder. In: 34th British Machine Vision Conference (BMVC) (2023)
18. Wald, T., Roy, S., Koehler, G., Disch, N., Rokuss, M.R., Holzschuh, J., Zimmerer, D., Maier-Hein, K.: SAM.MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model. In: Medical Imaging with Deep Learning (MIDL), short paper track (2023)
19. Wang, J., Xia, B.: Bounding Box Tightness Prior for Weakly Supervised Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 526–536 (2021)
20. Wu, Q., Zhang, Y., Elbatel, M.: Self-prompting Large Vision Models for Few-Shot Medical Image Segmentation. In: Domain Adaptation and Representation Transfer (MICCAI-DART) (2023)
21. Zhang, L., Deng, X., Lu, Y.: Segment Anything Model (SAM) for Medical Image Segmentation: A Preliminary Review. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 4187–4194 (2023)
22. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize Segment Anything Model with One Shot. In: International Conference on Learning Representations (ICLR) (2024)