This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SegAnyPET: Universal Promptable Segmentation from Positron Emission Tomography Images

Yichi Zhang[1,2] Le Xue[1,2] Wenbo Zhang[1,2] Lanlan Li[1] Yuchen Liu[1,2] Chen Jiang[2]
Yuan Cheng[1,2] Yuan Qi[1,2]

[1] Fudan University [2] Shanghai Academy of Artificial Intelligence for Science

Abstract

Positron Emission Tomography (PET) is a powerful molecular imaging tool that plays a crucial role in modern medical diagnostics by visualizing radio-tracer distribution to reveal physiological processes. Accurate organ segmentation from PET images is essential for comprehensive multi-systemic analysis of interactions between different organs and pathologies. Existing segmentation methods are limited by insufficient annotation data and varying levels of annotation, resulting in weak generalization ability and difficulty in clinical application. Recent developments in segmentation foundation models have shown superior versatility across diverse segmentation tasks. Despite the efforts of medical adaptations, these works primarily focus on structural medical images with detailed physiological structural information and exhibit limited generalization performance on molecular PET imaging. In this paper, we collect and construct PETS-5k, the largest PET segmentation dataset to date, comprising 5,731 three-dimensional whole-body PET images and encompassing over 1.3M 2D images. Based on the established dataset, we develop SegAnyPET, a modality-specific 3D foundation model for universal promptable segmentation from PET images. To issue the challenge of discrepant annotation quality, we adopt a cross prompting confident learning (CPCL) strategy with an uncertainty-guided self-rectification process to robustly learn segmentation from high-quality labeled data and low-quality noisy labeled data for promptable segmentation. Experimental results demonstrate that SegAnyPET can segment seen and unseen target organs using only one or a few prompt points, outperforming state-of-the-art foundation models and task-specific fully supervised models with higher accuracy and strong generalization ability for universal segmentation. As the first foundation model for PET images, we believe that SegAnyPET will advance the applications to various downstream tasks for molecular imaging. [1]

[1]The model and code will be publicly available at the project page.

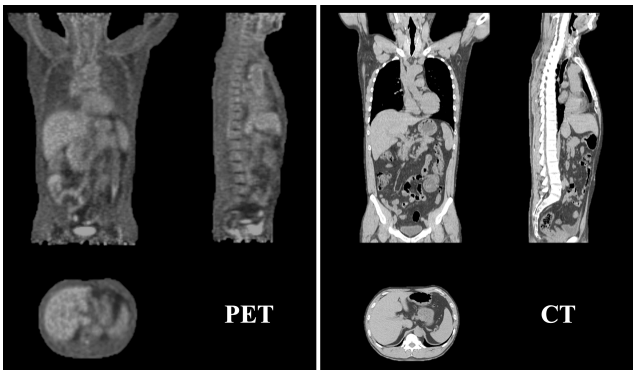


Figure 1. Visual comparison of an example case between molecular PET image and structural CT image.

1. Introduction

Medical imaging technology has become an indispensable tool in modern clinical diagnosis [13, 18]. The steady development of increasingly sophisticated imaging techniques has led to a dramatic increase in patient treatment outcomes, especially when diseases can only be observed internally within the patient's body [17]. In addition to commonly used medical modalities such as computed tomography (CT) and magnetic resonance imaging (MRI), which can provide detailed physiological structural information, positron emission tomography (PET) has also been widely applied in medical examinations as an emerging molecular imaging modality [29]. The purpose of PET imaging is to reveal the ongoing metabolic processes within the body of a patient. The radioactive decay of injected tracer emits positrons, leading to gamma photons, which are produced by annihilation and captured by detectors surrounding the patient to create a 3D map of metabolic activity. As the most widely used tracer, fluorodeoxyglucose (FDG) is used to assess local glucose uptake and evaluate organ metabolism [27] and presence of tumor metastasis [4], so as to monitor progress during the treatment process. Accurate organ segmentation from PET images is essential for comprehen-

sive multi-systemic analysis of interactions between different organs and pathologies [31, 33, 43]. However, due to the low resolution nature and the partial volume effects, PET images exhibit lower image quality and have less distinct boundaries compared to other structural images like CT (see Fig. 1), which presents a significant challenge for accurate segmentation.

With the unprecedented developments in deep learning, deep neural networks have been widely applied and achieved great success in many medical image segmentation tasks [9, 28]. However, the availability of PET images and annotations is far behind that of CT and MRI [10, 19]. A significant obstacle lies in the high costs associated with acquisition and annotation of PET images, and the varying levels of annotation due to the low image quality and the presence of partial volume effects. Although there are a few publicly available PET datasets [4, 25], they only focus on limited targets and do not cover these clinically important anatomic structures. In addition, existing deep models are often tailored for specific tasks, which require task-specific training data for model development and lack the capacity for generalization across different tasks. When adapted to new segmentation task, additional training with annotated data is required.

The emergence of pre-trained foundation models has sparked a new era due to their remarkable generalization abilities across a wide range of downstream tasks [23, 35]. For image segmentation, the introduction of the Segment Anything Model (SAM) [14] has gained massive attention as a promptable foundation model capable of generating fine-grade segmentation masks to unseen targets using positional prompts like points or bounding boxes. However, the original SAM is trained on natural images characterized by strong edge information, which differs significantly from medical images. When directly applying SAM without any adaptation, its performance significantly left behind task-specific segmentation models [22, 41]. Despite the attempts of fine-tuning and adapting SAM for medical images [5, 20], these works mainly focus on modalities like CT and MRI with detailed physiological structural information, while struggling to generalize to molecular PET imaging.

To bridge this gap, we collect and construct a large-scale PET dataset named **PETS-5k** with voxel-level multi-organ annotation. The dataset consists of 5,731 three-dimensional whole-body PET images, encompassing over 1.3M 2D images, which is the largest PET dataset to date. Based on the empirical analysis, we found that existing foundation models show poor generalization performance on PET images. To enable universal promptable segmentation from PET images, we present **SegAnyPET**, a modality-specific foundation model that can be efficiently and robustly adapted for universal promptable segmentation from PET images. Other than the original 2D architecture of SAM, SegAnyPET utilizes a 3D architecture to fully utilize the inter-slice context information of 3D PET volumes. To issue the challenge of discrepant annotation quality, we adopt a cross prompting confident learning (CPCL) strategy with an uncertainty-guided self-rectification process to efficiently learn from dataset with varying annotation quality for model training. The strategy does not require network modification and can be easily adapted to other promptable foundation models. Extensive experiments demonstrate that SegAnyPET can correctly segment both seen and unseen targets using only one or few prompt points and outperform state-of-the-art foundation models and task-specific fully supervised models by a large margin, underpinning its general-purpose segmentation ability. Our contribution can be summarized as follows:

- The up-to-date largest 3D PET segmentation dataset **PETS-5k**, with 5,731 three-dimensional PET images, encompassing over 1.3M 2D images, that are significantly larger than existing datasets.
- A modality-specific foundational model **SegAnyPET** for universal promptable segmentation from PET images, demonstrating superior performance compared with state-of-the-art methods and strong generalization ability.
- A cross prompting confident learning (CPCL) strategy for noisy-robust learning to efficiently learn from high-quality labeled data and low-quality noisy labeled data.

2. Related Work

Vision Foundation Models. Foundation models represent a rapidly expanding field in artificial intelligence research, focusing on developing large-scale, general-purpose models with capabilities applicable across various domains and applications. Pioneering vision foundation models have been primarily based on pre-training methods such as CLIP [26] and ALIGN [11] leverage contrastive learning techniques to train both text and image encoders. However, these models primarily excel in tasks that involve mapping images to text, such as classification. For foundation model of image segmentation, the Segment Anything Model (SAM) [14] represents a new paradigm of image segmentation with universal segmentation ability across different tasks and demonstrates strong potential in addressing a wide range of downstream tasks such as medical image analysis [23, 38].

Foundation Models for Medical Image Segmentation. As a crucial branch of image segmentation, recent studies have explored the generalization ability of SAM to biomedical image segmentation tasks of different modalities and targets [7, 21, 22], while some research also focuses on fine-tuning SAM [20] or adapting SAM-like foundation models [3, 32, 36] on large-scale medical datasets. These studies indicate that SAM has limited generalization ability when directly applied to medical image segmentation due to the significant distribution gap, and fine-tuning SAM or train-

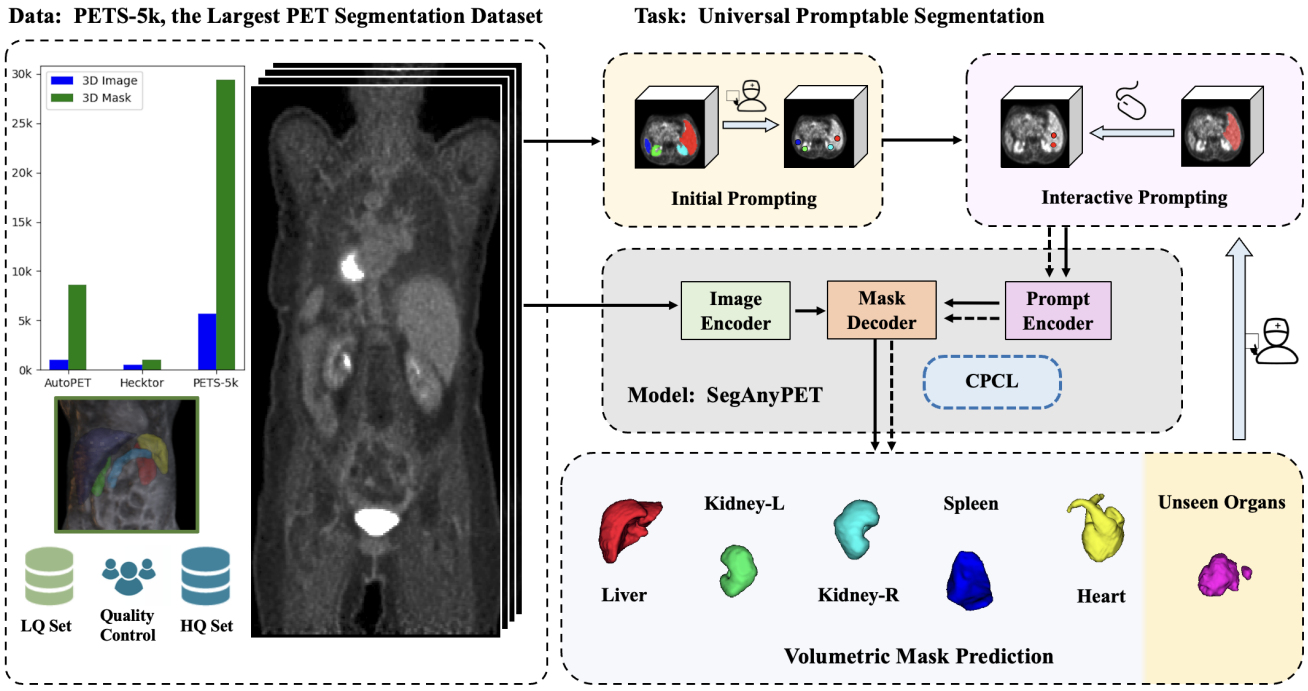


Figure 2. An overview of our work for PET segmentation foundation model. Firstly, we collect and construct PETS-5k, a large-scale PET segmentation dataset for developing the foundation model. Based on the principle of promptable segmentation, we develop SegAnyPET, a modality-specific 3D segmentation foundation model that can be efficiently adapted for universal segmentation of any target organs or lesions based on positional prompting from 3D PET images.

ing promptable segmentation foundation models on medical datasets can improve the unsatisfactory performance [41]. Other than developing promptable foundation models for medical imaging, another line of research aims to extend the promptable interactive segmentation manner into end-to-end fully automatic segmentation by auto-prompting [15, 39] or synergy in training supervised segmentation models [16, 42]. However, due to the scarcity of public datasets, existing adaptions mostly focus on structural medical images, while ignoring molecular PET images.

Annotation-Efficient Learning. Despite the outstanding performance in many medical image segmentation tasks [9, 34], the success of existing deep neural networks heavily relies on the massive radiologist-examined labeled data. Obtaining pixel-wise annotation is time-consuming and expertise-demanding, which motivates annotation-efficient learning strategies to train deep models with limited annotations [2, 12, 30]. Despite recent efforts of large-scale medical datasets [1, 37], these datasets are the integration of existing public datasets with different annotation quality. While some labels may be noisy with mistakes and could potentially hinder the learning procedure, it is important to develop noise-robust learning strategies to issue this challenge.

3. Methodology

In this paper, we aim to develop a modality-specific foundation model for universal segmentation from PET images. An overview of our work is shown in Fig. 2. Building upon the proposed large-scale PETS-5k dataset, we design a 3D promptable segmentation framework and introduce a cross prompting confident learning strategy to facilitate model training with different annotation quality. The details are elaborated in the following sections.

3.1. Dataset Construction

To develop a foundation model for universal segmentation from PET images, we collect and construct a large-scale PET dataset with 5,731 three-dimensional whole-body PET images, which is the largest PET dataset to date. In contrast to existing PET datasets with limited annotated targets, our work focuses on addressing a broader spectrum of multi-organ segmentation from whole-body images to facilitate representation learning for universal segmentation. All images are preliminarily annotated by AI and junior annotators to form the LQ set. A subset is then checked and refined by senior experts with quality control, serving as the HQ set. More information of the dataset are shown in the Appendix.

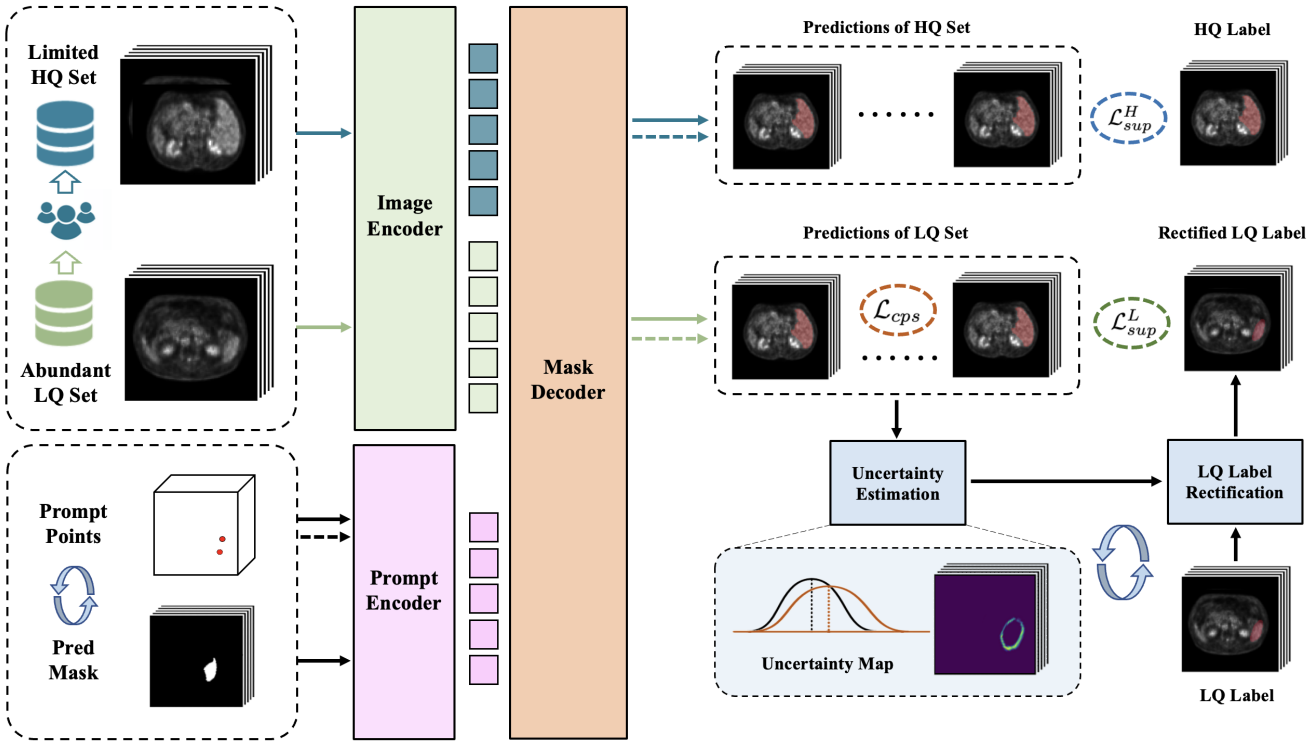


Figure 3. Illustration of the proposed cross prompting confident learning (CPCL) strategy for developing promptable segmentation foundation model based on both high quality and low quality annotations.

3.2. Network Architecture

Motivated by the recent advance of segmentation foundation models for natural images [14], we reformulate a holistic 3D structure in SegAnyPET to capture the spatial information directly from volumetric images for universal segmentation. Specifically, SegAnyPET consists of three main components, including an image encoder $\mathcal{F}_{IE}$, a prompt encoder $\mathcal{F}_{PE}$, and a mask decoder $\mathcal{F}_{D}$. The image encoder aims to transform input images into discrete embeddings, and the prompt encoder converts input prompts into compact embeddings by combining fixed positional encoding and adaptable prompt-specific embeddings. After that, the mask decoder receives the extracted information from both the image encoder and the prompt encoder, which is then up-sampled and passes through a multi-layer perception to generate the output segmentation masks. The processed feature map is then up-sampled and passes through a multi-layer perception to generate the output segmentation masks. Details of the network components are shown in the Appendix. Given an input image X and spatial prompt p, the output segmentation mask $\hat{Y}$ can be formally expressed as follows:

$$\hat{Y}_i = \begin{cases} \mathbf{O}_{H\times W\times D}, & i = 0 \\ \mathcal{F}_D(\mathcal{F}_{IE}(X), \mathcal{F}_{PE}(p_i, \hat{Y}_{i-1})), & i \geq 1 \end{cases} \tag{1}$$

where $\mathbf{O}_{H\times W\times D}$ is an all-zero matrix as the pre-defined initial segmentation mask. p_i represents the i_{th} prompt point in the training loop, where new point prompts are generated from error regions based on the previous segmentation mask to simulate the manual interactive segmentation workflow.

3.3. Cross Prompting Confident Learning

Due to the dependency of domain knowledge for annotating medical images, acquiring large amounts of high-quality labeled data is infeasible. A prevalent approach to mitigating this issue is collecting additional labeled data with varying annotation qualities like crowdsourcing from non-experts or using model-generated labels. Given that some annotations may be fraught with noise or inaccuracies, which could impede the learning process, it is imperative to devise noise-robust learning strategies to efficiently learn from both HQ Set containing expert-examined high-quality annotations and LQ Set containing low-quality annotations

Method	Prompt	Organ Segmentation DSC performance [%]					
		Liver	Kidney-L	Kidney-R	Heart	Spleen	Avg
SAM [14]	N points	26.55	9.38	9.10	14.44	6.30	13.15
MedSAM [20]	N points	0.25	0.19	1.32	0.27	0.27	0.46
SAM-Med3D [32]	1 point	51.63	21.01	19.17	60.11	25.41	35.46
SAM-Med3D-organ	1 point	80.25	44.70	35.76	74.00	69.23	60.79
SAM-Med3D-turbo	1 point	79.46	66.95	72.81	73.03	68.19	72.09
SegAnyPET	1 point	93.06	89.84	90.61	88.29	90.67	90.49
SAM [14]	3N points	43.85	23.21	22.16	29.09	11.83	26.03
MedSAM [20]	3N points	26.59	28.86	28.98	18.82	32.96	27.24
SAM-Med3D [32]	3 points	62.15	28.21	31.19	61.44	27.07	42.01
SAM-Med3D-organ	3 points	84.82	47.33	48.57	75.85	74.60	66.23
SAM-Med3D-turbo	3 points	84.11	74.05	76.17	75.24	73.34	76.58
SegAnyPET	3 points	93.36	90.25	90.95	88.86	91.10	90.90
SAM [14]	5N points	54.49	47.16	37.42	42.19	18.79	40.01
MedSAM [20]	5N points	36.53	37.53	39.22	24.71	41.30	35.86
SAM-Med3D [32]	5 points	61.05	31.05	31.98	61.88	29.75	43.14
SAM-Med3D-organ	5 points	85.52	49.56	54.40	76.30	75.13	68.18
SAM-Med3D-turbo	5 points	85.56	76.74	78.08	76.16	75.20	78.35
SegAnyPET	5 points	93.42	90.39	91.24	88.95	91.22	91.05
SegResNet (HQ) [24]	Auto	92.28	85.56	82.94	88.21	86.55	87.11
SwinUNETR (HQ) [6]	Auto	93.03	89.23	83.62	89.03	87.93	88.57
nnUNet (HQ) [9]	Auto	93.16	90.12	86.60	90.96	88.32	89.83
STUNet (HQ) [8]	Auto	93.15	90.11	85.60	90.38	88.83	89.61
SegResNet (All) [24]	Auto	91.43	86.29	85.73	88.88	86.45	87.76
SwinUNETR (All) [6]	Auto	91.15	84.98	84.32	88.13	84.89	86.69
nnUNet (All) [9]	Auto	93.74	88.21	88.26	90.50	88.30	89.80
STUNet (All) [8]	Auto	93.45	87.97	87.67	90.21	88.06	89.47

Table 1. Comparison of DSC performance with state-of-the-art general-propose foundation models for **interactive segmentation** and trainable task-specific models for **automatic segmentation** from PET images. N denotes the count of slices containing the target object (N ranges from 20 to 50 in our task).

with possible mistakes. This challenge is particularly pronounced for PET images due to the presence of partial volume artifacts and low image quality, and directly learning from these low-quality annotations may provide undesirable guidance. To issue the challenge of discrepant annotation quality, we adopt a cross prompting confident learning (CPCL) strategy. Figure 3 depicts the workflow of CPCL. The utilization of LQ set can be mainly divided into the following two parts.

Exploit Knowledge Inside the Images. Given that LQ labels could be the encumbrance for model training, it is natural to adopt the semi-supervised learning setting that casts LQ labels away and exploits the knowledge inside the image data alone. For promptable segmentation settings, the output heavily relies on the input prompts, and different prompts may cause variances in the segmentation results even when they refer to the same object given the same image. This phenomenon underscores the imperative to bolster the invariance of segmentation outputs across different prompts, which could be utilized as unsupervised regularization for training. Based on this motivation, we harness the capabilities inherent to the prompting mechanism and develop a cross prompting consistency strategy to learn from unlabeled data. Within each training loop, we estimate the unsupervised regularization loss as follows:

$$\bar{Y} = \frac{1}{n_{pt}} \sum_{i=1}^{n_{pt}} \hat{Y}_i \tag{2}$$

Algorithm 1 The training procedure of SegAnyPET.

Input: Training data $\mathcal{D}_H = \{X_{(i)}, Y_{H(i)}\}_{i=1}^{M}$ with high-quality annotations, and $\mathcal{D}_L = \{X_{(i)}, Y_{L(i)}\}_{i=1}^{N}$ with low-quality annotations

Output: SegAnyPET model parameters: Image Encoder $\mathcal{F}_{IE}$, Prompt Encoder $\mathcal{F}_{PE}$, Mask Decoder $\mathcal{F}_D$

Initialize training details
for $t \Leftarrow 1$ **to** t_{max} **do**
  Extract image embeddings $f_{img} = \mathcal{F}_{IE}(X)$
  Initialize $\hat{Y}_0$ as all zero matrix $\mathbf{O}$.
  for $i \Leftarrow 1$ **to** n_{pt} **do**
    Input prompt $p_i \Leftarrow \text{PromptGenerate}(Y, \hat{Y}_{i-1})$
    Update mask $\hat{Y}_i = \mathcal{F}_D(f_{img}, \mathcal{F}_{PE}(p_i, \hat{Y}_{i-1}))$
  end for
  Average prediction $\bar{Y} = \frac{1}{n_{pt}} \sum_{i=1}^{n_{pt}} \hat{Y}_i$
  Estimate uncertainty $U = -\sum p(\bar{Y}|X) \log p(\bar{Y}|X)$
  if X in $\mathcal{D}_H$ **then**
    Supervised loss $\mathcal{L}_{seg}^{H} \Leftarrow (\hat{Y}_i, Y)$
  else if X in $\mathcal{D}_L$ **then**
    Label rectification $\tilde{Y} \Leftarrow (Y, U)$
    Supervised loss $\mathcal{L}_{seg}^{L} \Leftarrow (\hat{Y}_i, Y, Y_{rec})$
    Regularization loss $\mathcal{L}_{cps} \Leftarrow (\bar{Y}, \{\hat{Y}_1, ..., \hat{Y}_{n_{pt}}\})$
  end if
  Total loss $\mathcal{L} = \mathcal{L}_{seg}^{H} + \lambda(\mathcal{L}_{cps} + \beta\mathcal{L}_{seg}^{L})$
  Update model parameters $(\mathcal{F}_{IE}, \mathcal{F}_{PE}, \mathcal{F}_D) \Leftarrow \mathcal{L}$
end for
return model

$$\mathcal{L}_{cps} = \sum_{i=1}^{n_{pt}} \|\bar{Y} - \hat{Y}_i\|^2 \tag{3}$$

where n_{pt} represents the setting of prompt point number in each training loop and $\bar{Y}$ represents the averaged segmentation result in the loop.

Self-Rectification Towards Effective Guidance. In addition to image-based information, effectively leveraging concomitant noisy labels inherent in LQ set is crucial to further improve the performance. To this end, we further propose an uncertainty-based self-rectification process to alleviate the potential misleading caused by label noises. Uncertainty estimation refers to the process of predicting the uncertainty or confidence of model predictions by assigning a probability value that represents the self-confidence associated with the prediction, so as to quantify the reliability of the output and to identify when the network may not be performing well. For image segmentation, the model assigns a probability value that represents the self-confidence associated with each pixel's prediction, and lower confidence (*i.e.* higher uncertainty) can be a heuristic likelihood of being mislabeled pixels.

Building upon the foundation of promptable segmentation, instead of adding perturbations to input images or models, we leverage the advantage of the prompting mechanism and approximate the segmentation uncertainty with different input prompts in the training loop. Following [40], we select the predictive entropy to approximate the uncertainty since it has a fixed range. The aleatoric uncertainty of a given image is formulated as follows:

$$U = -\sum p(\bar{Y}|X) \log p(\bar{Y}|X) \tag{4}$$

To exploit informative knowledge from LQ labels, we perform an uncertainty-based self-rectification process for noisy label refinement. With the guidance of estimated uncertainty, we filter out relatively unreliable regions with higher uncertainty as potential error maps. Then the rectification operation is formulated as:

$$\tilde{y}_i = y_i + \mathbb{I}(u_i > H) \cdot (-1)^{y_i} \tag{5}$$

where $\mathbb{I}(.)$ is the indicator function, y_i and u_i are the initial annotation and estimated uncertainty at i-th voxel. Then the rectified label of the whole volume $\tilde{Y}$ is $\{\tilde{y}\} \in \{0, 1\}_{H \times W \times D}$. For LQ set, we utilize rectified label for supervised training instead of the original noisy label. With the help of the label rectification process, the negative effects brought by label noises can be eliminated to some extent and provide more rewarding guidance.

3.4. Overall Training Procedure

Algorithm 1 presents the overall training procedure of SegAnyPET. The total loss for model training is weighted combination of the supervised loss $\mathcal{L}_{seg}^{H}$ on HQ set, the unsupervised regularization loss $\mathcal{L}_{cps}$ on LQ set, and the rectified supervised loss $\mathcal{L}_{seg}^{L}$ on LQ set, calculated by:

$$\mathcal{L} = \mathcal{L}_{seg}^{H} + \lambda(\mathcal{L}_{cps} + \beta\mathcal{L}_{seg}^{L}) \tag{6}$$

where β is a fixed weighting coefficient, and λ is a ramp-up trade-off weighting coefficient to avoid domination by misleading targets at the early training stage.

4. Experiment & Results

4.1. Experimental Setup

The proposed PETS-5k dataset consists of 5,731 three-dimensional whole-body PET images is used for training and internal validation. Among the dataset, a subset of 100 images is with high-quality annotations verified by domain experts, which is split into 40 cases serving as HQ training set and 60 cases serving as internal test set. The remaining images with possible mislabeled or unlabeled pixels serving as the LQ training set. To evaluate the generalization ability for universal segmentation, we annotate 7 additional training invisible organs from the internal test set. We further

Method	Prompt	Training Invisible Organ Segmentation DSC Performance [%]							
		Aorta	Lung-LL	Lung-LR	Lung-UL	Lung-UR	Lung-MR	Prostate	Avg
SAM [14]	N points	2.55	9.60	12.24	11.17	15.23	9.75	8.01	9.79
MedSAM [20]	N points	0.10	0.89	0.70	0.97	1.70	2.57	0.51	1.06
SAM-Med3D [32]	1 point	11.29	25.60	45.95	37.94	47.15	31.64	9.93	29.93
SAM-Med3D-organ	1 point	24.12	45.71	42.16	56.32	61.71	37.85	32.25	42.87
SAM-Med3D-turbo	1 point	21.28	31.55	37.82	44.36	53.56	39.34	25.46	36.20
SegAnyPET	1 point	82.99	88.32	89.76	89.47	91.91	89.17	91.67	89.04
SAM [14]	3N points	6.45	13.91	16.88	16.40	22.04	15.45	19.30	15.78
MedSAM [20]	3N points	17.95	24.85	20.18	18.29	21.23	23.86	25.82	21.74
SAM-Med3D [32]	3 points	13.38	29.77	46.62	45.46	50.50	36.93	13.73	33.77
SAM-Med3D-organ	3 points	28.76	48.77	50.06	64.17	69.10	51.05	33.84	49.39
SAM-Med3D-turbo	3 points	27.37	36.55	47.00	57.91	66.17	57.66	31.21	46.27
SegAnyPET	3 points	84.51	89.09	90.54	90.26	92.52	89.72	92.19	89.83
SAM [14]	5N points	14.58	19.37	23.37	22.97	31.62	24.88	41.03	25.40
MedSAM [20]	5N points	21.73	27.45	25.55	23.05	27.91	28.24	37.69	27.37
SAM-Med3D [32]	5 points	14.07	37.54	48.46	50.68	51.01	38.43	13.37	36.23
SAM-Med3D-organ	5 points	29.95	53.12	54.02	65.87	71.35	52.55	35.25	51.73
SAM-Med3D-turbo	5 points	31.92	42.05	48.36	64.34	69.65	61.15	32.77	50.03
SegAnyPET	5 points	84.63	89.21	90.69	90.34	92.63	89.80	92.57	89.98

Table 2. Generalization performance to **training invisible organs** with comparison to state-of-the-art state-of-the-art general-propose foundation models for interactive segmentation from PET images.

annotate a small subset of 100 cases with all 12 testing organs from the AutoPET dataset [4] for external validation, entitled AutoPET-Organ. Following the setting for promptable segmentation, the input prompts are simulated based on the ground truth mask with random perturbations. We compare proposed SegAnyPET with state-of-the-art segmentation foundation models [14, 20, 32] and training-based task-specific models [6, 8, 9, 24]. More information of the implementation details are shown in the Appendix.

4.2. Experimental Results

Comparison with Universal Foundation Models. The original SAM [14] and MedSAM [20] are designed for 2D segmentation tasks and cannot handle 3D inputs directly, necessitating input prompts for each 2D slice containing the target to conduct slice-by-slice segmentation. In contrast, SegAnyPET and other 3D SAM medical adaptions [32] can be directly utilized to segment the target organs from input volume with one or few prompts. Table 1 presents the performance of SegAnyPET with state-of-the-art foundation models on our internal test set under different prompt settings. We can observe that 2D models exhibit lower performance due to the lack of volumetric spatial information. In contrast, 3D models can achieve better performance with less manual prompting effort. However, all these existing foundation models show poor generalization performance on PET segmentation task. Compared with state-of-the-art methods, SegAnyPET exhibits significantly better segmentation performance and outperforms the second-ranked method by a significant improvement up to 18.40% on average DSC. Besides, we visualize the segmentation results of SegAnyPET and other segmentation foundation models for organ segmentation in Fig. 4, which intuitively show that SegAnyPET performs more precise segmentation with higher DSC using the same or fewer manual efforts.

Comparison with Task-Specific Models. In addition to foundation models, we also compare with state-of-the-art task-specific models in Tab. 1. These task-specific models are trained for the segmentation of target organs. When adapting to training invisible organs, additional annotation and training are required. We conduct two experimental settings, including training on HQ data alone and training with the combination of HQ and LQ data. Our observations show that incorporating additional LQ data does not consistently enhance model performance and may even degrade the performance in some scenarios. Despite the efforts of prompting, SegAnyPET demonstrates superior performance while retaining generalization capability to unseen targets.

Generalization to Unseen Organs. To validate the generalization performance of SegAnyPET, we conduct

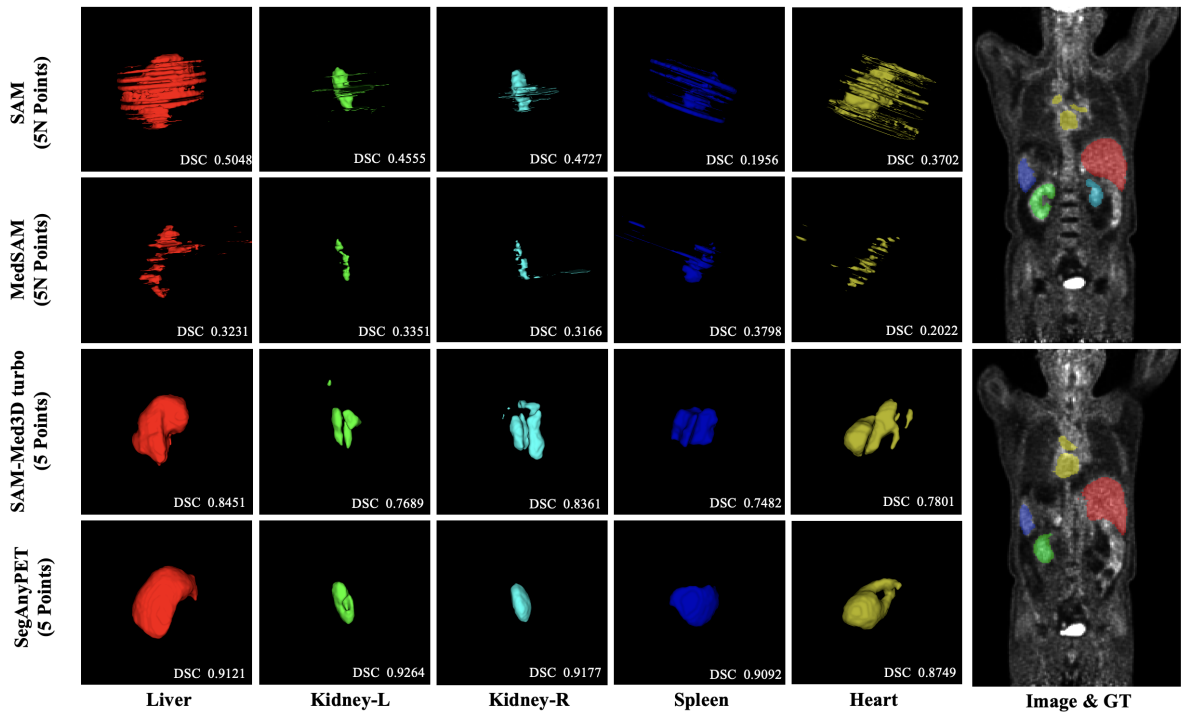


Figure 4. Visual comparison of segmentation results of different promptable segmentation foundation models for PET organ segmentation.

additional experiments of foundation models for promptable segmentation of training invisible organs. As shown in Tab. 2, we observe SegAnyPET also exhibits strong generalization performance and outperforms state-of-the-art segmentation foundation models. The experimental results demonstrate that by efficiently learning representations from large-scale PET images, SegAnyPET exhibits comparable performance for training invisible organs. We further evaluate SegAnyPET on training invisible AutoPET and AutoPET-Organ dataset with significant distribution gap compared with PETS-5k dataset.

Ablation Analysis. We conduct ablation experiments to evaluate the effectiveness of proposed training strategy. Three different training strategies are conducted including fine-tuning with mixed HQ and LQ, training with additional unsupervised consistency regularization, and the proposed CPCL strategy. Experimental results show that the proposed strategy can achieve better performance on both seen and unseen targets. More detailed experimental results and analysis are shown in the Appendix.

5. Discussion & Conclusion

In this work, we propose SegAnyPET, a modality-specific foundation model for universal promptable segmentation from 3D PET images. To this end, we collect and construct a large-scale PET segmentation dataset, PETS-5k, which consists of 5,731 three-dimensional whole-body PET images, encompassing over 1.3M 2D images to train the foundation model. To facilitate the model training with different annotation quality, we design a cross prompting confident learning strategy building on the foundation of promptable segmentation for noise-robust learning. Our model shows substantial capabilities in segmenting both training visible and invisible target organs using only one or few prompt points. Its performance significantly exceeds both state-of-the-art segmentation foundation models and task-specific fully supervised models.

While SegAnyPET boasts strong capabilities, there are still inherent limitations due to manual efforts for promptable segmentation, which could potentially reduce the usability of the model compared with fully automatic segmentation. We aim to incorporate semantic information together with positional prompts as additional support to enable efficient and precise auto-prompting for automatic segmentation in the future. We also aim to enlarge existing data and task to more comprehensive scenarios including different tracers and tumor lesion segmentation. In conclusion, as the first foundation model for PET images, we believe that SegAnyPET will advance the developments of universal segmentation and serve as a powerful pre-trained model to facilitate the applications to various downstream tasks for molecular imaging.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 82394432 and 92249302), and the Shanghai Municipal Science and Technology Major Project (Grant No. 2023SHZDZX02). The computations in this research were performed using the CFFF platform of Fudan University.

References

[1] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024. 3

[2] Veronika Cheplygina, Marleen De Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019. 3

[3] Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *Advances in neural Information Processing Systems*, 37:110746–110783., 2024. 2

[4] Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022. 1, 2, 7

[5] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, page 103324, 2024. 2

[6] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021. 5, 7

[7] 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? *Medical Image Analysis*, 92:103061, 2024. 2

[8] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023. 5, 7

[9] 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. 2, 3, 5, 7

[10] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 2

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[12] Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, and Cheng Jin. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, page 107840, 2023. 3

[13] Hany Kasban, MAM El-Bendary, and DH Salama. A comparative study of medical imaging techniques. *International Journal of Information Science and Intelligent System*, 4(2): 37–58, 2015. 1

[14] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 4, 5, 7

[15] Wenhui Lei, Wei Xu, Kang Li, Xiaofan Zhang, and Shaoting Zhang. Medlsam: Localize and segment anything model for 3d ct images. *Medical Image Analysis*, 99:103370, 2025. 3

[16] Ning Li, Lianjin Xiong, Wei Qiu, Yudong Pan, Yiqian Luo, and Yangsong Zhang. Segment anything model for semi-supervised medical image segmentation via selecting reliable pseudo-labels. In *International Conference on Neural Information Processing*, pages 138–149. Springer, 2023. 3

[17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[18] Charles J. Lynch and Conor Liston. New machine-learning technologies for computer-aided diagnosis. *Nature Medicine*, 24:1304–1305, 2018. 1

[19] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2022. 2

[20] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:1–9, 2024. 2, 5, 7

[21] Jun Ma, Sumin Kim, Feifei Li, Mohammed Baharoon, Reza Asakereh, Hongwei Lyu, and Bo Wang. Segment anything in medical images and videos: Benchmark and deployment. *arXiv preprint arXiv:2408.03322*, 2024. 2

[22] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 2

[23] Michael Moor, Oishi Banerjee, Zahra F H Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Ra-

jpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023. 2

[24] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 311–320. Springer, 2019. 5, 7

[25] Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, et al. Head and neck tumor segmentation in pet/ct: the hecktor challenge. *Medical image analysis*, 77:102336, 2022. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[27] Silin Ren, Priscille Laub, Yihuan Lu, Mika Naganawa, and Richard E Carson. Atlas-based multiorgan segmentation for dynamic abdominal pet. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 4(1):50–62, 2019. 1

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[29] Johannes Schwenck, Dominik Sonanini, Jonathan M Cotton, Hans-Georg Rammensee, Christian la Fougère, Lars Zender, and Bernd J Pichler. Advances in pet imaging of cancer. *Nature Reviews Cancer*, 23(7):474–490, 2023. 1

[30] Yuyan Shi, Jialu Ma, Jin Yang, Shasha Wang, and Yichi Zhang. Beyond pixel-wise supervision for medical image segmentation: From traditional models to foundation models. *arXiv preprint arXiv:2404.13239*, 2024. 3

[31] Tao Sun, Zhenguo Wang, Yaping Wu, Fengyun Gu, Xiaochen Li, Yan Bai, Chushu Shen, Zhanli Hu, Dong Liang, Xin Liu, et al. Identifying the individual metabolic abnormities from a systemic perspective using whole-body pet imaging. *European Journal of Nuclear Medicine and Molecular Imaging*, 49(8):2994–3004, 2022. 2

[32] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, et al. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. *European Conference on Computer Vision*, 2024. 2, 5, 7

[33] Hanzhong Wang, Xiaoya Qiao, Wenxiang Ding, Gaoyu Chen, Ying Miao, Rui Guo, Xiaohua Zhu, Zhaoping Cheng, Jiehua Xu, Biao Li, et al. Robust and generalizable artificial intelligence for multi-organ segmentation in ultra-low-dose total-body pet imaging: a multi-center and cross-tracer study. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–15, 2025. 2

[34] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023. 3

[35] Martin J. Willemink, Holger R. Roth, and Veit Sandfort. Toward foundational deep learning models for medical imaging in the new era of transformer networks. *Radiology. Artificial intelligence*, 46:210284, 2022. 2

[36] Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In *European Conference on Computer Vision*, pages 207–229. Springer, 2024. 2

[37] Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023. 3

[38] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91:102996, 2024. 2

[39] Yichi Zhang, Shiyao Hu, Chen Jiang, Yuan Cheng, and Yuan Qi. Segment anything model with uncertainty rectification for auto-prompting medical image segmentation. *arXiv preprint arXiv:2311.10529*, 2023. 3

[40] Yichi Zhang, Rushi Jiao, Qingcheng Liao, Dongyang Li, and Jicong Zhang. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. *Artificial Intelligence in Medicine*, 138:102476, 2023. 6

[41] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171:108238, 2024. 2, 3

[42] Yichi Zhang, Jin Yang, Yuchen Liu, Yuan Cheng, and Yuan Qi. Semisam: Enhancing semi-supervised medical image segmentation via sam-assisted consistency regularization. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3982–3986. IEEE, 2024. 3

[43] Liqin Zhu, David Finkelstein, Culian Gao, Lei Shi, Yongdong Wang, Dolores López-Terrada, Kasper Wang, Sarah Utley, Stanley Pounds, Geoffrey Neale, et al. Multi-organ mapping of cancer risk. *Cell*, 166(5):1132–1146, 2016. 2