P^2SAM: Probabilistically Prompted SAMs Are Efficient Segmentator for Ambiguous Medical Images

Yuzhi Huang*
School of Informatics,
Xiamen University, China
yzhuang13@stu.xmu.edu.cn

Chenxin Li*†
The Chinese University of Hong Kong, Hong Kong SAR.
chenxinli@link.cuhk.edu.hk

Zixu Lin
School of Informatics,
Xiamen University, China
zxlin@stu.xmu.edu.cn

Hengyu Liu
The Chinese University of Hong Kong, Hong Kong SAR.
piang_lhy@163.com

Haote Xu
School of Informatics,
Xiamen University, China
hotxu@stu.xmu.edu.cn

Yifan Liu
The Chinese University of Hong Kong, Hong Kong SAR.
1155195605@link.cuhk.edu.hk

Yue Huang†
School of Informatics,
Xiamen University, China
yhuang2010@xmu.edu.cn

Xinghao Ding
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Xiamen University, China
School of Informatics,
Xiamen University, China
dxh@xmu.edu.cn

Xiaotong Tu
School of Informatics,
Xiamen University, China
xttu@xmu.edu.cn

Yixuan Yuan
The Chinese University of Hong Kong, Hong Kong SAR.
yxyuan@ee.cuhk.edu.hk

Abstract

Generating diverse plausible outputs from a single input is crucial for addressing visual ambiguities, exemplified in medical imaging where experts may provide varying semantic segmentation annotations for the same image. Existing methods handles ambiguous segmentation relying on probabilistic modeling and extensive multi-output annotated data while often struggles with limited ambiguously labeled datasets common in real-world applications. To surmount the challenge, we propose P^2SAM, a novel framework that leverages the Segment Anything Model (SAM)'s prior knowledge for ambiguous object segmentation. By transforming SAM's sensitivity to prompts into an advantage, we introduce a prior probabilistic space for prompts. Experimental results show that P^2SAMsignificantly enhances medical segmentation precision and diversity using minimal ambiguously annotated samples. Benchmarking against state-of-the-art methods demonstrates superior performance with just 5.5% of the training data (+12% D_{max}). This approach marks a significant advancement towards deploying probabilistic models in data-limited real-world scenarios. Website: https://p2-sam.github.io/.

∗ Equal contribution † Corresponding author.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680628

CCS Concepts

• **Computing methodologies → Image segmentation**.

Keywords

Probabilistic modeling, Medical image segmentation, Prompting for foundation model

ACM Reference Format:
Yuzhi Huang*, Chenxin Li*†, Zixu Lin, Hengyu Liu, Haote Xu, Yifan Liu, Yue Huang†, Xinghao Ding, Xiaotong Tu, and Yixuan Yuan. 2024. P^2SAM: Probabilistically Prompted SAMs Are Efficient Segmentator for Ambiguous Medical Images. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680628

1 Introduction

Numerous complex situations in the physical domain often encompass a spectrum of potentially viable solutions for multiple purposes [1, 19, 20, 22, 26]. This is particularly noticeable in medical imaging analysis [7, 8, 29, 52, 53, 65, 76], and the relevant surgical application [11, 23, 27, 44, 51], where inherent ambiguity in boundary structures and multiple plausible annotations arise due to limitations in imaging mechanisms, indeterminate boundaries

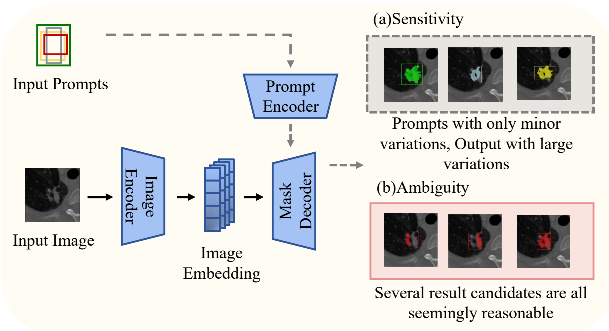

Figure 1: The challenges of SAM in deterministic segmentation: a) High prompt sensitivity: SAM produces significantly diverse segmentation results with subtle variations in the input prompt box. b) Output ambiguity: For a given prompt, SAM can generate multiple reasonable segmentation results, especially when segmenting images with complex hierarchical structures.

among medical professionals, and varying experiences [39]. The task paradigm associated with such ambiguity in the data itself, defined as "single input, multiple outputs", is referred to as Ambiguous Segmentation [19]. The advantages of this paradigm are evident. For instance, providing multiple regions of a lesion automatically can assist doctors in focusing on the areas of concern, rather than on ambiguous regions. However, traditional models, which establish a one-to-one mapping between inputs and outputs [28, 30, 68], generating a unique segmentation map for each image, are fundamentally incapable of addressing such *ambiguous* scenarios.

To tackle the *ambiguity* issue, a variety of works have restructured the conventional "one-to-one" segmentation process by amalgamating insights from multiple experts and producing a range of outputs that account for pixel uncertainty and diverse image annotations. For instance, the Probabilistic U-Net [19], which combines U-net [62] and cVAE [8, 17, 67], effectively encapsulates annotation distributions to generate an array of segmentation maps. Concurrently, models such as PHiSeg [3], PixelSeg [83], and CIMD [59] tackle uncertainty via varied sampling and introduce new accuracy metrics for segmentation.

Despite these advances, current methods still struggle to balance segmentation fidelity and diversity. This is primarily attributed to the fact that these probabilistic modeling techniques often sacrifice prediction accuracy to increase the complexity of distribution space and generate more diverse annotations. Additionally, compared to learning conventional deterministic mappings, probabilistic modeling inherently requires more training samples to fit an underlying "one-to-many" distribution of uncertainty. However, in actual clinical diagnoses, there is often a shortage of high-quality lesion samples annotated by multiple experts, leading to suboptimal probabilistic modeling. To alleviate this degradation of performance in practical applications, this paper presents a pioneering study for probabilistic and ambiguous representation learning under data-limited settings [24, 37, 40, 56].

Our intuition derives from the recent progress of Visual Foundation Models (VFM) [10, 58, 70], specifically SAM [18], which have been pre-trained on more than a billion masks across eleven million natural images. These models leverage prior knowledge extracted during large-scale pre-training to facilitate the segmentation of downstream tasks. Despite SAM's impressive generalization capabilities in segmentation, certain constraints have been observed: (1) SAM exhibits notable sensitivity to minor variations in prompts, necessitating precision in prompt inputs as illustrated in Figure 1(a), where minor translational or scaling operations on the input detection box prompt can cause significant alterations in SAM's output. (2) SAM may grapple with the issue of prompt ambiguity when confronted with target objects possessing complex hierarchical structures. This is due to the challenges in defining boundaries of elements at different levels, all of which seem feasible for a given prompt, as shown in Figure1(b). This dilemma requests SAM to generate multiple deterministic segmentation mask candidates.

In this work, we explore an unconventional perspective in an attempt to turn the *ambiguity* disadvantage of SAM in a deterministic segmentation task into an advantage in probabilistic ambiguous segmentation. Specifically, we aspire to address the following two pertinent questions. First, considering the sensitivity of SAM's output to prompts, how can we model the distribution of prompts to control the generation of diversified segmentation outputs? Second, given the ambiguity of SAM's output to prompts, how can we modulate these ambiguous outputs as a reference for probabilistic modeling segmentation?

Inspired by these insights, we propose a Probabilistically Prompted Segment Anything framework, dubbed as P^2SAM, which adeptly addresses ambiguous segmentation tasks in medical imaging. Specifically, we initially design a prompt generation network that automatically generate plausible prompts without require the manual prompt design. This network probabilistically models the representation of input prompts in SAM, thereby simulating the moderating effect of different prompts on the output results. After sampling from the aforementioned prompt distribution, we employ a diversity-aware ambiguous ensemble algorithm to adaptively perceive the optimal ensemble weights of diverse ambiguous segmentation outputs, further modulating the multiple ambiguous segmentation masks generated by SAM. Lastly, by integrating the aforementioned strategies into the efficient fine-tuning of the current off-the-line medical SAM framework, our framework demonstrates powerful data efficiency in carrying out ambiguous segmentation tasks. Our contributions can be summarized as:

- We introduce a Probabilistically Prompted Segment Anything Model (P^2SAM), leveraging SAM's powerful segmentation prior and untapped latent uncertainty knowledge.
- We architect a generative network to model the prior distribution of prompts, conditioned on input images, generating meaningful prompt distributions for SAM.
- We develop a diversity-aware ambiguous ensemble algorithm, guiding the model to adaptively weigh SAM's different masks, enhancing segmentation diversity.
- Extensive empirical benchmarking shows our method outperforms state-of-the-art in both accuracy and diversity

of segmentation, approaching physician-level performance while using significantly fewer training samples.

2 Related Work

Ambiguous Image Segmentation. Ambiguous image segmentation methodologies aim to encapsulate the aleatoric uncertainties and inherent unpredictability of labels employed for segmentation. A plethora of research has proposed diverse techniques to quantify aleatoric uncertainty. Preliminary research focused on enhancing a conventional U-Net[5, 14, 50, 63, 75] with a probabilistic component to generate multiple predictions for an identical image, typically achieved by incorporating a conditional variational autoencoder (cVAE) [66]. The cVAE's low-dimensional latent space encodes potential segmentation variants. In [19], samples from this latent space are upscaled and concatenated at the U-Net's final layer. Numerous methodologies extend this setup to a hierarchical variant [3, 20, 84]. Other research utilizes normalizing flows to allow for a distribution more expressive than the Gaussian distribution in the cVAE [64, 69], switch to a discrete latent space [57], or incorporate variational dropout and directly use inter-grader variability as a training target [13]. Several other methods do not rely on the Probabilistic U-Net [8, 16, 31, 55, 77]. Monteiro *et al.* [55] propose a network utilizing a low-rank multivariate normal distribution to model the logit distribution. Kassapis *et al.* [16] leverage adversarial training to learn potential label maps based on the logits of a trained segmentation network. Zhang *et al.* [83] employ an autoregressive PixelCNN to model the conditional distribution between pixels. Lastly, Gao *et al.* [9] use a mixture of stochastic experts, where each expert network estimates a mode of uncertainty [38], and a gating network predicts the probabilities that an input image is segmented by one of the experts. Different from previous efforts, our methodology is the inaugural exploration of employing large-scale pre-trained models for ambiguous image segmentation.

Prompting Segmentation Foundation Models. In recent years, the potential of large-scale vision models for many tasks, such as image segmentation and image restoration [41, 42, 73**?** , 74], has been demonstrated by several concurrent works, inspired by language foundation models [4, 21, 45, 47, 78, 82]. These Segmentation Foundation Models (SFMs) like the Segment Anything Model (SAM) [18] and SEEM [86], have showcased impressive segmentation performance across diverse downstream datasets. SAM, utilizing a data engine with a model-in-the-loop annotation [48, 49], learns a promptable segmentation framework that generalizes to downstream scenarios in a zero-shot manner. Other models like Painter [71] and SegGPT [72] introduce a robust in-context learning paradigm and can segment any images given an image-mask prompt. SEEM [86], on the other hand, presents a general segmentation model prompted by multi-modal references, such as language and audio, incorporating versatile semantic knowledge. These advances in SFMs, largely driven by the *promptable segmentation* design, involve two types of prompts: semantic prompts (e.g., free-form texts) and spatial prompts (e.g., points or bounding boxes) [18, 33, 34, 46, 79, 86]. Despite these advances, acquiring suitable prompts for SFMs remains a largely under-explored area. Instead, this work aims to investigate the generation of effective prompts for SAM, with a focus on utilizing pre-training knowledge to complete ambiguous image segmentation.

3 Method

3.1 A Revisit of Segment Anything Model (SAM)

Segment Anything Model (SAM), an exemplar of transformer-based architecture, has demonstrated remarkable efficacy in the realms of natural language processing and image recognition tasks. Specifically, SAM employs a vision transformer-based image encoder to extract salient image features, prompt encoders to assimilate user interactions, and subsequently, a mask decoder to generate segmentation results and confidence scores, contingent on the image embedding, prompt embedding, and output token. SAM is a tripartite structure comprising of a prompt encoder, an image encoder, and a lightweight mask decoder, denoted respectively as Enc_P, Enc_I, and Dec_M. As an interactive framework, SAM ingests an image I, and a set of prompts P, which may be a point, a box, or a coarse mask. Specifically, SAM first employs Enc_I to obtain the input image feature, and adopts Enc_P to encode the human-given prompts of a length k into prompt tokens as follows

$$F_I = \mathrm{Enc}_I(I), \quad T_P = \mathrm{Enc}_P(P), \tag{1}$$

where $F_I \in \mathbb{R}^{h \times w \times c}$ and $T_P \in \mathbb{R}^{k \times c}$, where the resolution of the image feature map is represented by h, w, and the feature dimension is denoted by c. Subsequently, the encoded image and prompts are introduced into the decoder Dec_M for interaction based on attention mechanisms. SAM constructs the decoder's input tokens by concatenating several learnable mask tokens T_M as prefixes to the prompt tokens T_P. These mask tokens are accountable for generating the mask output, formulated as follows

$$M = \mathrm{Dec}_M\left(F_I,\ \mathrm{Concat}(T_M, T_P)\right), \tag{2}$$

where M denotes the final segmentation mask predicted by SAM.

3.2 Lifting SAM to Probabilistic Space

Ambiguous segmentation tasks require multiple segmentation results for a single input to more accurately reflect the true distribution of real-world scenarios. Interestingly, we observe an inherent ambiguity in SAM, where minor positional modifications to prompts lead to substantial alterations in SAM's segmentation output. This observation catalyzes our consideration for probabilistic modeling of prompt variations. By utilizing a distribution of prompt embedding, rather than a single deterministic prompt, we can effectively modulate the model output, as

$$\tilde{T_P} \sim \mathcal{P}_{PE}(\theta), \tag{3}$$

where $\mathcal{P}_{PE}$ denotes a probability distribution for the space of prompt embedding, $\tilde{T_P}$ is specific a prompt sampling from the given distribution at one time. Formally, by implementing multiple rounds of sampling, we can construct a probabilistic mapping of segmentation outputs with respect to their prompts, formulated as the format of expectation

$$\mathbb{E}_{\tilde{M} \sim \mathcal{P}_M(\vartheta)} = \mathbb{E}_{\tilde{T_P} \sim \mathcal{P}_{PE}(\theta)} \mathrm{Dec}_M\left(F_I,\ \mathrm{Concat}(T_M, \tilde{T_P})\right) \tag{4}$$

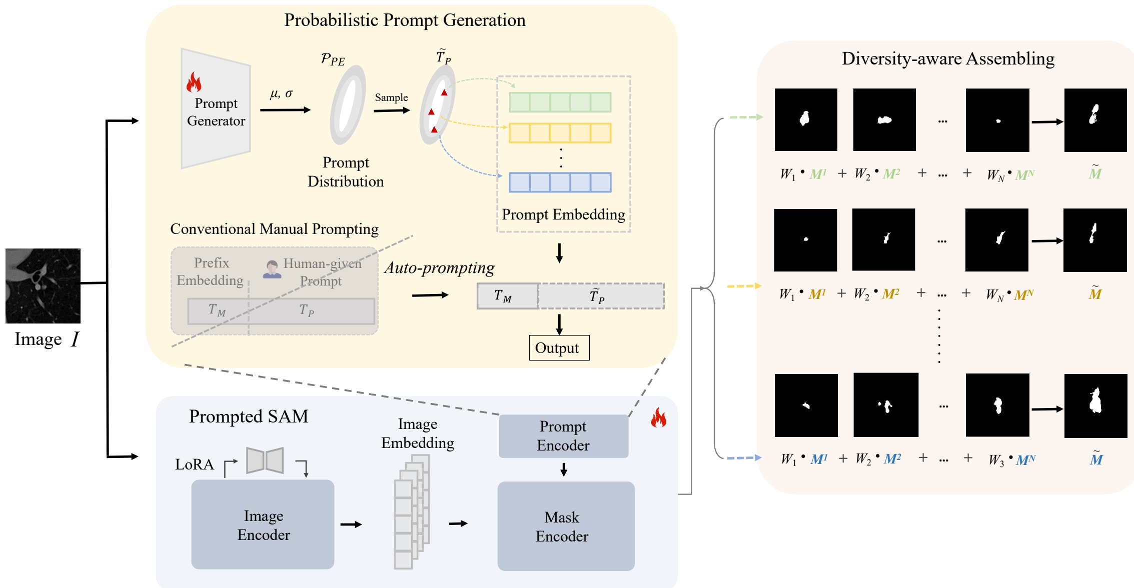

Figure 2: P²SAM Training Pipeline. We first lift the conventional SAM prompting to the probabilistic space, by leveraging a network targeting at generating prompt distribution. Then we sample the prompt embedding from the probabilistic latent space and instill it into SAM to unlock the capacity of SAM in "one-to-many" ambiguous segmentation. We carefully design a diversity-aware assembling that perceives the inherent diversity in SAM and turn it to ensembled ambiguous output.

where $\tilde{M}$ denotes the SAM output corresponding to the prompt sampling every time, which can also be interpreted as the sampling from a virtual distribution $\mathcal{P}_M$ for the segmentation results which obeys the parameters ϑ. As a result, we can construct an optimized probability distribution $\tilde{T_P} \sim \mathcal{P}_{PE}(\theta)$ by narrowing the gap between $\tilde{M} \sim \mathcal{P}_M(\vartheta)$ and the ground-truth distribution.

3.3 Instance-conditional Probabilistic Prompt Generation

To model the probability distribution of prompt embedding, it is imperative to estimate the parameters θ of this distribution. We adopt an axisymmetric Gaussian distribution to characterize the prompt embedding, which is dictated by two crucial parameters: μ (mean) and σ (standard deviation). To accurately model the prompt embedding, we have designed a dedicated prompt generation network. This network comprises two primary components: an encoder and an Axis-Gaussian generation network. The encoder, composed of several simple convolution blocks, is designed to extract image features. Subsequently, these feature maps are introduced to the Axis-Gaussian generation network, which is a convolutional network structure [85]. Then we can sample a prompt embedding from the given Gaussian distribution by

$$\tilde{T_P} \sim \mathcal{N}(\mu_I, \mathrm{diag}(\sigma_I)), \tag{5}$$

where μ_I and σ_I denotes the parameters characterized for image I.

We further dedicated a Prompt Generation Network, made up of several straightforward convolutional blocks, aims to extract features from the image. Subsequently, these feature maps are fed into the Axis-Gaussian generation network, which is also a convolutional network structure [85]. Considered the variation in salient regions within an image suggests that the required prompt location and size should also differ, making it impractical to apply a uniform probability distribution model to prompt embeddings. Hence, we introduce image prior knowledge into the Prompt Generation Network during forward inference. By incorporating this prior knowledge, the network can customize a unique axis Gaussian distribution for each image I, thus achieving more precise sampling for the prompt embedding, as

$$\mu_I, \sigma_I = PGN(\theta|I), \tag{6}$$

where PGN stands for the Prompt Generation Network modeled by the parameters ω, and μ and σ respectively denote the mean and standard deviation of the Axis Gaussian Distribution generated by the network, where $\mu, \sigma \in \mathbb{R}^N$ with N=256.

3.4 Diversity-aware Assembling

We assume that SAM implicitly models the probability of default adaptive prompts in Section 3.2, but how does this adaptive prompt focus on images with complex hierarchical structures? When dealing with images with complex hierarchical structures, it is not clear how this adaptive prompt effectively focuses on key areas. Especially when SAM faces segmentation tasks on an image containing multiple salient targets, it often faces the challenge of segmentation ambiguity. To overcome this ambiguity, SAM generates multiple segmentation masks to segment the salient regions of the image from different levels and perspectives. Although this method can provide multi angle segmentation results, these results often fail to fully reflect the certainty and uniqueness of segmentation, making it difficult to provide a more convincing and clear segmentation.

To integrate the multi-scale segmentation masks of SAM under ambiguous prompts, we introduce a ambiguous integration strategy with diverse sensitivities. This strategy relies on SAM to obtain multi-scale outputs, which refer to the original segmentation results of multi-scale output by SAM, as $\{M^1, M^2, ..., M^N\}$, where N denotes the number of scale. On top of this, we adopt learnable mask weights $\mathcal{W} = \{w_1, w_2, ..., w_N\} \in \mathbb{R}^N$, and calculate final mask output through weighted summation as:

$$\tilde{M} = \Sigma_{i=1}^{N} w_i * \tilde{M}^i \tag{7}$$

In order to learn the optimal weights, we fine-tune SAM and also trained this parameter. By adopting this strategy, we can effectively learn and understand the scale perception of objects while preserving the deep knowledge of the pre-trained model. In addition, it can adaptively integrate masks of multiple scales to achieve precise output of the optimal segmentation scale for the target object.

3.5 Overall Optimization Procedure

During the optimization of the entire framework, we found that the direct application of SAM is limited in certain specific vertical scenarios. Therefore, we propose to fine-tune SAM to our tasks first, followed by efficient probabilistic prompt training. Specifically, the overall optimization process is divided into two crucial stages. In the first stage, we aim to fine-tune the key modules within the SAM model to empower its adaptation ability [25, 32, 35, 36], enhancing the generalization to applicable domains [6, 43, 61, 80], including the modulation module, which integrates diverse outputs, the mask decoder, the prompt encoder, and the image encoder. Notably, we fine-tune the image encoder via the Low Rank Adaptation (LoRA) [12] strategy, thereby keeping the original parameters of the image encoder unchanged. Specially, at the data level, we only use non-empty labels for model fine-tuning and training, which speeds up the model's adaptation. The loss function used in this process is as follows:

$$\mathcal{L}_1(\tilde{M}, \tilde{GT}; Enc_P, Dec_M, \mathcal{W}, Enc_I(LoRA)) = \ell_{seg}(\tilde{M}, \tilde{GT}), \tag{8}$$

At the second stage, upon establishing the benchmark performance for the segmentation task and the capability to handle ambiguous sets, we further enhanced our model to address the challenges associated with ambiguous segmentation. In this stage, we froze the parameters for all components and concentrated on training the prompt generation network. The image is provided as input to the prompt generation network, which is responsible for generating a precise prompt probability distribution. Following this, a series of approximate yet distinct prompts are obtained by sampling from this distribution. These prompts are then fed into the SAM model to achieve ambiguous segmentation. During this process, we employed the following loss function:

$$\mathcal{L}_2(\tilde{M}, \tilde{GT}; PGN) = \ell_{seg}(\tilde{M}, \tilde{GT}). \tag{9}$$

4 Experiment

4.1 Datasets

Lung lesion segmentation (LIDC-IDRI). This dataset is publicly accessible and comprises a substantial collection of 1018 lung Computed Tomography (CT) scans, derived from 1010 distinct subjects. This dataset is notable for its inclusion of manual annotations, contributed by a panel of four domain experts. This feature makes the dataset a robust and accurate reflection of the typical ambiguity often encountered in CT imaging, as referenced in the study [2]. A diverse group of 12 radiologists lent their expertise to provide annotation masks for this dataset, further enhancing its value. The version of the dataset we use in this study is the one obtained after the second reading. In this phase, the domain experts were presented with the annotations made by other radiologists. This process allowed them to make new adjustments based on the feedback and insights of their peers, thereby ensuring the dataset's annotations are comprehensive, accurate, and reflective of a broad spectrum of expert opinions.

Brain tumour segmentation in 3D (BraTS 2017). The BraTS 2017 [15] dataset encompasses 285 cases of 3D MRI images, each comprising 155 slices. Every slice is provided in four modalities (T1, T1ce, T2, and Flair) and has been meticulously annotated across four classes by expert radiologists: background (BG), non-enhanced/necrotic tumor core (NET), oedema (OD), and enhanced tumor core (ET). We overlay and amalgamate annotations from these various categories, transforming the results into a binary mask that solely includes the foreground and background. This procedure is designed to generate multiple segmentation masks to mimic actual ambiguous segmentation scenarios, thereby enhancing the rigor and reliability of the experiment.

4.2 Implementation Details

In our experiments, we employed SAMed [81] as our primary segmentation network, a specialized variant of SAM designed for automatic medical image segmentation without the need for manually designed prompts. We utilized two datasets: LIDC and BraTS 2017, each partitioned into training, validation, and testing sets in a 60:20:20 ratio. For the LIDC dataset, it comprises 1,018 lung CT scans from 1,010 patients, and the associated four annotations for lung nodules from different experienced radiologists. The images are resized to 128x128 pixels. For the BraTS 2017 dataset, we used the T1 modality and resized the images to 128x128 pixels. We utilized the dataset's three levels of lesion intensity - non-enhanced/necrotic tumour core (NET), non-enhanced/necrotic tumour core (NET)+ oedema (OD), and non-enhanced/necrotic tumour core (NET)+ oedema (OD) + enhanced tumour core (ET) - as three ambiguous segmentation labels for each slice. Our model training process consisted of two stages. Initially, we fine-tuned only the SAM and the learnable weights in the diversity-aware assembling module, using the Adam optimizer with a learning rate of 1e-3 for 100 epochs. Subsequently, we froze the SAM parameters and the learnable weight modules, focusing solely on training the prompt generator network with an adjusted learning rate of 1e-5. This two-stage approach allowed us to effectively leverage the pre-trained knowledge in SAM while optimizing our model for ambiguous segmentation tasks.

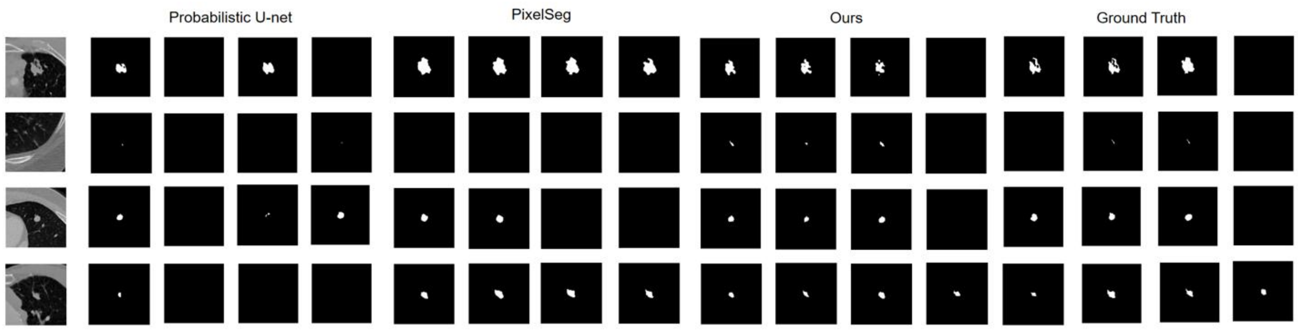

Figure 3: Comparative qualitative analysis with the advanced methods, including Probabilistic U-net [19] and PixelSeg [83]. Examples of available four ground-truth expert labels and sampled segmentation masks are provided.

Table 1: Comparison of GED, HM-IoU, and D_{max} quantitative results of LIDC(5.5% of the entire dataset) using state-of-the-art ambiguous segmentation networks.

Method	LIDC (500 samples)			
	GED16(↓)	GED32(↓)	HM-IoU(↑)	D_{max}(↑)
Probabilistic U-net [19]	0.325	0.337	0.324	0.251
CAR [16]	0.8849	0.905	0.179	0.567
PixelSeg [83]	0.328	0.299	0.495	<u>0.731</u>
Mose [9]	<u>0.290</u>	<u>0.276</u>	<u>0.510</u>	0.652
P^{2}SAM (Ours)	**0.208**	**0.206**	**0.627**	**0.919**

4.3 Evaluation Metrics

Generalized Energy Distance (GED). A common metric for ambiguous image segmentation that leverages distance between observations by comparing the distribution of segmentation [19], as

$$D^2_{GED}(P_{gt}, P_{out}) = 2\mathbb{E}[d(S, Y)] - \mathbb{E}[d(S, S')] - \mathbb{E}[d(Y, Y')], \quad (10)$$

where, d corresponds to the distance measure $d(x, y) = 1-IoU(x, y)$, Y and Y' are independent samples of P_{gt} and S and S' are sampled from P_{out}. Lower energy indicates better agreement between prediction and the ground truth distribution of segmentations.

Maximum Dice Matching (D_{max}). In medical diagnosis cases, empty sets, which indicate no abnormalities are also valid diagnoses. However, in this case, the Dice metric will be undefined, hence we set Dice = 1 in those cases. Specially, the Dice score is defined as:

$$Dice(\hat{Y}, Y) = \begin{cases} \frac{2|Y \cap \hat{Y}|}{|Y|+|\hat{Y}|}, & \text{if } Y \cup \hat{Y} \neq \emptyset \\ 1, & \text{otherwise.} \end{cases} \quad (11)$$

To calculate the best prediction accuracy for a set of prediction samples, we calculated the Dice score between each prediction result and each ground truth. We define the set of all Dice scores $\mathbf{D}_i$ for each individual ground truth Y_i, as follows

$$\mathbf{D}_{max} = max\{Dice(\hat{Y}_1, Y_i), Dice(\hat{Y}_2, Y_i), ..., Dice(\hat{Y}_N, Y_i)\}, \quad (12)$$

where $\mathbf{D}_i$ is a collection of Dice scores calculated between each ground truth Y_i and all the provided predictions, and $\mathbf{D}_{max}$ is the maximum result among all $\mathbf{D}_i$.

Hungarian-Matched Intersection over Union (HM-IoU). GED excessively rewards sample diversity, which cannot reflect the sufficiency of the sample. Therefore, the Hungarian Matching IoU (HM-IoU) is proposed to calculate the optimal 1:1 between annotation and prediction, which better represents the fidelity of the sample. The Hungarian algorithm finds the optimal 1:1 match between objects in two sets, for which we use $IoU(Y, Y')$ to determine the similarity between the two samples.

4.4 Results on LIDC-IDRI

We compare our approach to numerous recent stochastic segmentation methods: Probabilistic U-Net [19], Hierarchical Probabilistic U-Net (HProb. U-net) [20], PhiSeg [3], Stochastic Segmentation Network (SSN) [54], Calibrated Adversarial Refinement (CAR) [16], PixelSeg [83], Mixture of Stochastic Experts (MoSE) [9], Collectively Intelligent Medical Diffusion (CIMD) [60], and SAMed [81]. Table 1 and Table 2 present the results. We used two data versions to train the model, which are 500 samples from the training set and all available training set samples. We annotate the metrics calculated with n samples using a subscript, *i.e.*, GED_n and $HM-IoU_n$, where n is set to the common values found in the compared literature. The results show that our method significantly outperforms other

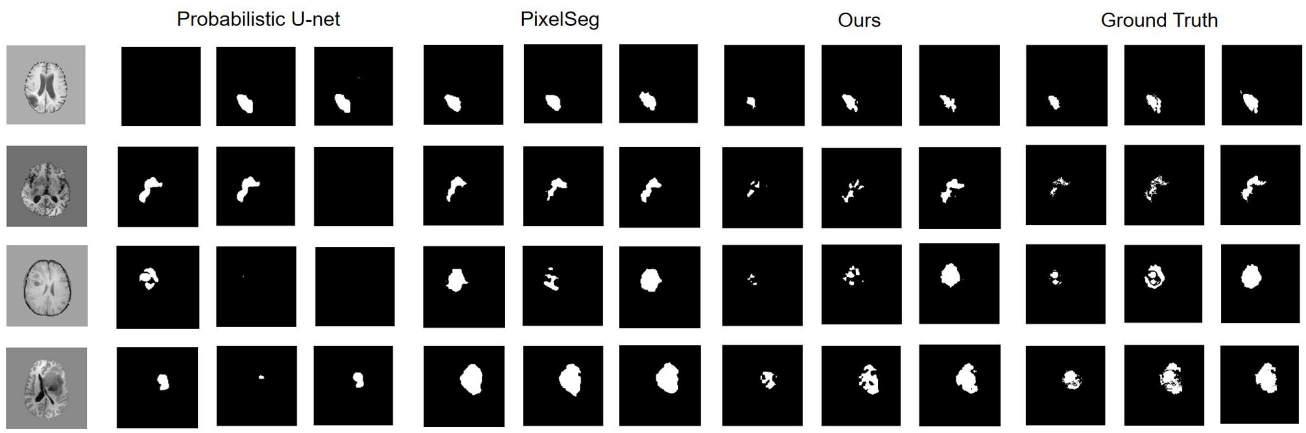

Figure 4: Comparative qualitative analysis with the advanced methods including Probabilistic U-net [19] and PixelSeg [83]. Examples of the available three ground-truth expert labels and sampled segmentation masks are provided.

Table 2: Comparison of GED, HM-IoU, and D_{max} quantitative results of LIDC (all samples) using state-of-the-art ambiguous segmentation networks.

Method	**LIDC (all samples)**			
	GED16(↓)	GED32(↓)	HM-IoU16(↑)	D_{max}(↑)
Probabilistic U-net [19]	0.324	0.303	0.423	0.370
HProb. U-net [20]	0.270	—	0.530	—
PHiseg [3]	0.262	0.247	0.595	—
SSN [54]	0.259	0.243	0.555	—
CAR [16]	0.252	—	0.549	0.732
PixelSeg [83]	0.243	0.245	0.614	0.814
CIMD [60]	0.234	0.218	0.587	—
Mose [9]	0.234	0.230	0.623	0.702
SAMed [81]	0.380	0.362	0.357	0.703
P^2SAM (Ours)	**0.218**	**0.216**	**0.679**	**0.933**

Table 3: Comparison of GED, HM-IoU, D_{max} and D_{mean} quantitative results of BraTS2017 using state-of-the-art ambiguous segmentation networks.

Method	**BraTS2017 (500 samples)**				**BraTS2017 (all samples)**			
	GED(↓)	HM-IoU(↑)	D_{max}(↑)	D_{mean}(↑)	GED(↓)	HM-IoU(↑)	D_{max}(↑)	D_{mean}(↑)
Probabilistic U-net [19]	0.154	0.427	0.517	0.346	**0.225**	0.521	0.645	0.464
PixelSeg [83]	0.549	0.414	0.516	**0.373**	0.419	0.528	0.785	**0.561**
SAMed [81]	0.189	0.216	0.407	0.355	0.267	0.411	0.716	0.432
P^2SAM (Ours)	**0.134**	**0.435**	**0.730**	0.363	0.238	**0.593**	**0.881**	0.494

state-of-the-art networks in various metrics on two different training sample datasets. Specifically, a higher D_{max} score indicates a high match between the distribution of the generated samples and the actual situation. Meanwhile, higher HM-IoU and lower GED scores comprehensively reflect the diversity and consistency of the samples, effectively quantifying the degree of agreement between prediction and annotation.

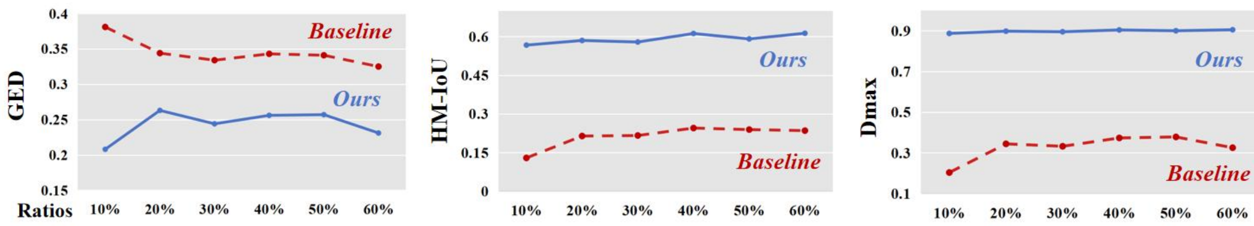

Figure 5: Comparison of GED (↓), HM-IoU (↑) and D_{max}(↑) between ours and baseline models under five different data ratios on LIDC-IDRI dataset.

Table 4: Ablation study of the key strategies of the proposed P²SAM on LIDC-IDRI dataset.

Method	GED(↓)	D_{max}(↑)	HM-IoU (↑)
Vanilla Adapted SAM	0.381	0.705	0.359
SAM + Probabilistic Prompt	<u>0.340</u>	0.803	<u>0.454</u>
SAM + Diversity Assembling	0.376	<u>0.853</u>	0.402
P²SAM (Full Model)	**0.208**	**0.919**	**0.627**

Evaluating ambiguous networks is challenging, and qualitative results can be a valuable indicator of performance, especially for complex cases. Figure 3 presents predictions from the test dataset for all models. P²SAM demonstrates visually superior and diverse results compared to previous state-of-the-art methods, particularly excelling in ultrasound modalities with minimal error. It successfully captures all lesions, including small structures, while maintaining diversity in segmentation masks. By injecting stochasticity at each hierarchical feature representation, P²SAM achieves diverse and accurate segmentation across all datasets.

4.5 Results on BraTS2017

Table 3 presents the quantitative results of P²SAM, PixelSeg [83], SAMed [81], and Probabilistic U-net [19] on the BraTS 2017 dataset. P²SAM demonstrates notable performance advantages, particularly in GED, D_{max}, and HM-IoU metrics, highlighting its effectiveness and robustness compared to the baseline methods. In terms of D_{mean}, P²SAM performs comparably to the baselines. Importantly, this performance is achieved with a limited training dataset, yet P²SAM generates a wider range of accurate segmentation samples. This outcome not only validates the efficiency of P²SAM but also underscores its practical value in addressing ambiguous medical image segmentation tasks.

Furthermore, the qualitative results illustrated in Figure 4 provide additional insights into the proposed method's performance relative to other techniques. The proposed method demonstrates superior accuracy and plausibility in segmentation results while also excelling in generating a more diverse range of predictions. This diversity is particularly valuable, offering a more comprehensive understanding and interpretation of the data, thus enhancing the overall effectiveness of the segmentation process.

4.6 Ablation Study

We conduct the ablation study for P²SAM on LIDC dataset, as shown in Table 4. **Vanilla Adapted SAM** denotes that we simply fine tune a SAM model for the multi-output task as a baseline. **SAM + Probabilistic Prompt** denotes that we introduced a prompt generator network on the fine tuned SAM to guide the generation of ambiguous prompts. Compared with the benchmark fine tuned SAM model, the introduction of PGN resulted in significant improvements in the three key performance indicators of GED, D_{max}, and HM-IoU, especially in the significant growth of GED. This result clearly indicates that the fusion of PGN not only enriches the output of the SAM model, but also significantly enhances the diversity of the output, further improving the performance of the model in ambiguous medical image segmentation tasks. **SAM + Diversity Assembling** means that we introduce learnable weights for diversity-aware assembling in our baseline model to guide the SAM model in outputting ambiguous segmentation results. Compared to the baseline model, there was little change in the GED indicator, but we observed significant improvement in the D_{max} and HM-IoU indicators. This result clearly indicates that by using learnable weights modules, we can effectively integrate the multiple outputs of the SAM model, thereby generating samples that are both representative and more accurate. Compared with all of the above variants, **Full Model** of (P²SAM) achieves the best results when all components work together. It appears that when any component is removed, the performance drops accordingly, revealing the effectiveness of our design.

4.7 Conclusion

This paper presents P^2SAM, tackling the inherent ambiguity prevalent in real-world visual scenarios, particularly in medical image segmentation. By leveraging the prior knowledge of the Segment Anything Model (SAM) and transforming its inherent drawback into an advantage, we demonstrate significant improvements in the precision and diversity of medical segmentation. Despite the challenges posed by limited availability of ambiguously annotated samples, our method outperforms state-of-the-art methods in rigorous benchmarking experiments, achieving superior segmentation precision and diversified outputs with fewer training data. This signifies a substantial step towards the practical deployment of probabilistic models in real-world scenarios with limited data.

5 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 82172033, Grant 52105126, Grant 82272071, and Grant 62271430; and in part by the Dreams Foundation of Jianghuai Advance Technology Center; and in part by the Open Fund of the National Key Laboratory of Infrared Detection Technologies.

References

[1] Sharib Ali, Noha Ghatwary, Debesh Jha, Ece Isik-Polat, Gorkem Polat, Chen Yang, Wuyang Li, Adrian Galdran, Miguel-Ángel González Ballester, Vajira Thambawita, et al. 2024. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *Sci. Rep.* (2024).

[2] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics* 38, 2 (2011), 915–931.

[3] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. 2019. Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 119–127.

[4] T.B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Askell Amanda, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Henighan Tom, Rewon Child, A. Ramesh, DanielM. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, EricJ. Sigler, Mateusz Litwin, Scott Gray, Chess Benjamin, Jack Clark, Christopher Berner, McCandlish Sam, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv: Computation and Language,arXiv: Computation and Language* (May 2020).

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*. Springer, 205–218.

[6] Chaoqi Chen, Luyao Tang, Feng Liu, Gangming Zhao, Yue Huang, and Yizhou Yu. 2022. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. *Advances in Neural Information Processing Systems* 35 (2022), 33302–33315.

[7] Zhen Chen, Wuyang Li, Xiaohan Xing, and Yixuan Yuan. 2023. Medical federated learning with joint graph purification for noisy label learning. *MedIA* (2023).

[8] Zhiyuan Ding, Qi Dong, Haote Xu, Chenxin Li, Xinghao Ding, and Yue Huang. 2022. Unsupervised Anomaly Segmentation for Brain Lesions Using Dual Semantic-Manifold Reconstruction. In *International Conference on Neural Information Processing*. Springer, 133–144.

[9] Zhitong Gao, Yucong Chen, Chuyu Zhang, and Xuming He. 2022. Modeling Multimodal Aleatoric Uncertainty in Segmentation with Mixture of Stochastic Expert. *arXiv preprint arXiv:2212.07328* (2022).

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[11] Zhibin He, Wuyang Li, Tuo Zhang, and Yixuan Yuan. 2023. H 2 GM: A Hierarchical Hypergraph Matching Framework for Brain Landmark Alignment. In *MICCAI*.

[12] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[13] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. 2019. Supervised uncertainty quantification for segmentation with multiple annotations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 137–145.

[14] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 1055–1059.

[15] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. 2018. *Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge*. 287–297. https://doi.org/10.1007/978-3-319-75238-9_25

[16] Elias Kassapis, Georgi Dikov, Deepak K Gupta, and Cedric Nugteren. 2021. Calibrated adversarial refinement for stochastic semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7057–7067.

[17] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, AlexanderC Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. [n. d.]. Segment Anything. ([n. d.]).

[19] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. 2018. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems* 31 (2018).

[20] Simon AA Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. 2019. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077* (2019).

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[22] Chenxin Li, Brandon Y Feng, Zhiwen Fan, Panwang Pan, and Zhangyang Wang. 2023. Steganerf: Embedding invisible information within neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 441–453.

[23] Chenxin Li, Brandon Y Feng, Yifan Liu, Hengyu Liu, Cheng Wang, Weihao Yu, and Yixuan Yuan. 2024. EndoSparse: Real-Time Sparse View Synthesis of Endoscopic Scenes using Gaussian Splatting. *arXiv preprint arXiv:2407.01029* (2024).

[24] Chenxin Li, Mingbao Lin, Zhiyuan Ding, Nie Lin, Yihong Zhuang, Yue Huang, Xinghao Ding, and Liujuan Cao. 2022. Knowledge condensation distillation. In *European Conference on Computer Vision*. Springer, 19–35.

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

[26] Chenxin Li, Hengyu Liu, Zhiwen Fan, Wuyang Li, Yifan Liu, Panwang Pan, and Yixuan Yuan. 2024. GaussianStego: A Generalizable Stenography Pipeline for Generative 3D Gaussians Splatting. *arXiv preprint arXiv:2407.01301* (2024).

[27] Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. 2024. Endora: Video Generation Models as Endoscopy Simulators. *arXiv preprint arXiv:2403.11050* (2024).

[28] Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, and Yixuan Yuan. 2024. U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. *arXiv preprint arXiv:2406.02918* (2024).

[29] Chenxin Li, Xinyu Liu, Cheng Wang, Yifan Liu, Weihao Yu, Jing Shao, and Yixuan Yuan. 2024. GTP-4o: Modality-prompted Heterogeneous Graph Learning for Omni-modal Biomedical Representation. *arXiv preprint arXiv:2407.05540* (2024).

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

[31] Chenxin Li, Yunlong Zhang, Jiongcheng Li, Yue Huang, and Xinghao Ding. 2021. Unsupervised anomaly segmentation using image-semantic cycle translation. *arXiv preprint arXiv:2103.09094* (2021).

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

[33] Wuyang Li, Zhen Chen, Baopu Li, Dingwen Zhang, and Yixuan Yuan. 2021. Htd: Heterogeneous task decoupling for two-stage object detection. *TIP* (2021).

[34] Wuyang Li, Xiaoqing Guo, and Yixuan Yuan. 2023. Novel Scenes & Classes: Towards Adaptive Open-set Object Detection. In *ICCV*. 15780–15790.

[35] Wuyang Li, Xinyu Liu, and Yixuan Yuan. 2022. Scan++: Enhanced semantic conditioned adaptation for domain adaptive object detection. *TMM* (2022).

[36] Wuyang Li, Xinyu Liu, and Yixuan Yuan. 2022. SIGMA: Semantic-complete Graph Matching for Domain Adaptive Object Detection. In *CVPR*.

[37] Wuyang Li, Xinyu Liu, and Yixuan Yuan. 2023. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *TPAMI* (2023).

[38] Wuyang Li, Xinyu Liu, and Yixuan Yuan. 2024. CLIFF: Continual Latent Diffusion for Open-Vocabulary Object Detection. In *ECCV*.

[39] Wuyang Li, Chen Yang, Jie Liu, Xinyu Liu, Xiaoqing Guo, and Yixuan Yuan. 2021. Joint polyp detection and segmentation with heterogeneous endoscopic data. In *ISBI Workshop: EndoCV 2021*.

[40] Zhehan Liang, Yu Rong, Chenxin Li, Yunlong Zhang, Yue Huang, Tingyang Xu, Xinghao Ding, and Junzhou Huang. 2021. Unsupervised large-scale social network alignment via cross network embedding. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1008–1017.

[41] Yunlong Lin, Zhenqi Fu, Ge Meng, Yingying Wang, Yuhang Dong, Linyu Fan, Hedeng Yu, and Xinghao Ding. 2023. Domain-irrelevant Feature Learning for Generalizable Pan-sharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3287–3296.

[42] Yunlong Lin, Tian Ye, Sixiang Chen, Zhenqi Fu, Yingying Wang, Wenhao Chai, Zhaohu Xing, Lei Zhu, and Xinghao Ding. 2024. AGLLDiff: Guiding Diffusion Models Towards Unsupervised Training-free Real-world Low-light Image Enhancement. *arXiv preprint arXiv:2407.14900* (2024).

[43] Yiyang Lin, Bowei Zeng, Yifeng Wang, Yang Chen, Zijie Fang, Jian Zhang, Xiangyang Ji, Haoqian Wang, and Yongbing Zhang. 2022. Unpaired multi-domain stain transfer for kidney histopathological images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1630–1637.

[44] Hengyu Liu, Yifan Liu, Chenxin Li, Wuyang Li, and Yixuan Yuan. 2024. LGS: A Light-weight 4D Gaussian Splatting for Efficient Surgical Scene Reconstruction. *arXiv preprint arXiv:2406.16073* (2024).

[45] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[46] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

[47] Xiaoyuan Liu, Wuyang Li, Takeshi Yamaguchi, Zihan Geng, Takuo Tanaka, Din Ping Tsai, and Mu Ku Chen. 2024. Stereo Vision Meta-Lens-Assisted Driving Vision. *ACS Photonics* (2024).

[48] Xinyu Liu, Wuyang Li, and Yixuan Yuan. 2022. Intervention & interaction federated abnormality detection with noisy clients. In *MICCAI*. 309–319.

[49] Xinyu Liu, Wuyang Li, and Yixuan Yuan. 2023. Decoupled Unbiased Teacher for Source-Free Domain Adaptive Medical Object Detection. *TNNLS* (2023).

[50] Xinyu Liu, Wuyang Li, and Yixuan Yuan. 2024. DiffRect: Latent Diffusion Label Rectification for Semi-supervised Medical Image Segmentation. *arXiv preprint arXiv:2407.09918* (2024).

[51] Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. 2024. EndoGaussian: Gaussian Splatting for Deformable Surgical Scene Reconstruction. *arXiv preprint arXiv:2401.12561* (2024).

[52] Yifan Liu, Wuyang Li, Jie Liu, Hui Chen, and Yixuan Yuan. 2023. GRAB-Net: Graph-based boundary-aware network for medical point cloud segmentation. *IEEE Transactions on Medical Imaging* (2023).

[53] Yifan Liu, Jie Liu, and Yixuan Yuan. 2022. Edge-oriented point-cloud transformer for 3D intracranial aneurysm segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 97–106.

[54] Miguel Monteiro, LoicLe Folgoc, DanielCoelhode Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Markvander Wilk, and Ben Glocker. 2020. Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty. *Cornell University - arXiv,Cornell University - arXiv* (Jun 2020).

[55] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. 2020. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems* 33 (2020), 12756–12767.

[56] Panwang Pan, Zhiwen Fan, Brandon Y Feng, Peihao Wang, Chenxin Li, and Zhangyang Wang. 2023. Learning to estimate 6dof pose from limited data: A few-shot, generalizable approach using rgb images. *arXiv preprint arXiv:2306.07598* (2023).

[57] Di Qiu and Lok Ming Lui. 2020. Modal Uncertainty Estimation via Discrete Latent Representation. *arXiv preprint arXiv:2007.12858* (2020).

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

[59] Aimon Rahman, JeyaMariaJose Valanarasu, Ilker Hacihaliloglu, and VishalM Patel. [n. d.]. Ambiguous Medical Image Segmentation using Diffusion Models. ([n. d.]).

[60] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. 2023. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11536–11546.

[61] Zhijie Rao, Jingcai Guo, Luyao Tang, Yue Huang, Xinghao Ding, and Song Guo. 2023. Srcd: Semantic reasoning with compound domains for single-domain generalized object detection. *arXiv preprint arXiv:2307.01750* (2023).

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

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

[64] Raghavendra Selvan, Frederik Faye, Jon Middleton, and Akshay Pai. 2020. Uncertainty quantification in medical image segmentation with normalizing flows. In *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11*. Springer, 80–90.

[65] Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep learning in medical image analysis. *Annual review of biomedical engineering* 19, 1 (2017), 221–248.

[66] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[67] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. 2015. Learning structured output representation using deep conditional generative models. *Neural Information Processing Systems,Neural Information Processing Systems* (Dec 2015).

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

[69] MM Amaan Valiuddin, Christiaan GA Viviers, Ruud JG van Sloun, Peter HN de With, and Fons van der Sommen. 2021. Improving Aleatoric Uncertainty Quantification in Multi-annotated Medical Image Segmentation with Normalizing Flows. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer, 75–88.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[71] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. 2022. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. (Dec 2022).

[72] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. [n. d.]. SegGPT: Segmenting Everything In Context. ([n. d.]).

[73] Yingying Wang, Xuanhua He, Yuhang Dong, Yunlong Lin, Yue Huang, and Xinghao Ding. 2024. Cross-Modality Interaction Network for Pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[74] Yingying Wang, Yunlong Lin, Ge Meng, Zhenqi Fu, Yuhang Dong, Linyu Fan, Hedeng Yu, Xinghao Ding, and Yue Huang. 2023. Learning high-frequency feature enhancement and alignment for pan-sharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*. 358–367.

[75] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. 2019. Nas-unet: Neural architecture search for medical image segmentation. *IEEE access* 7 (2019), 44247–44257.

[76] Haipeng Xu, Chenxin Li, Longfeng Zhang, Zhiyuan Ding, Tao Lu, and Huihua Hu. 2024. Immunotherapy efficacy prediction through a feature re-calibrated 2.5 D neural network. *Computer Methods and Programs in Biomedicine* 249 (2024), 108135.

[77] Haote Xu, Yunlong Zhang, Liyan Sun, Chenxin Li, Yue Huang, and Xinghao Ding. 2022. AFSC: Adaptive Fourier Space Compression for Anomaly Detection. *arXiv preprint arXiv:2204.07963* (2022).

[78] Qiushi Yang, Wuyang Li, Baopu Li, and Yixuan Yuan. 2023. MRM: Masked Relation Modeling for Medical Image Pre-Training with Genetics. In *ICCV*. 21452–21462.

[79] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. 2023. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790* (2023).

[80] Bowei Zeng, Yiyang Lin, Yifeng Wang, Yang Chen, Jiuyang Dong, Xi Li, and Yongbing Zhang. 2022. Semi-supervised pr virtual staining for breast histopathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 232–241.

[81] Kaidong Zhang and Dong Liu. 2023. Customized Segment Anything Model for Medical Image Segmentation. (Apr 2023).

[82] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. [n. d.]. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. ([n. d.]).

[83] Wei Zhang, Xiaohong Zhang, Sheng Huang, Yuting Lu, and Kun Wang. 2022. PixelSeg: Pixel-by-Pixel Stochastic Semantic Segmentation for Ambiguous Medical Images. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4742–4750.

[84] Wei Zhang, Xiaohong Zhang, Sheng Huang, Yuting Lu, and Kun Wang. 2022. A Probabilistic Model for Controlling Diversity and Accuracy of Ambiguous Medical Image Segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4751–4759.

[85] Yunlong Zhang, Chenxin Li, Xin Lin, Liyan Sun, Yihong Zhuang, Yue Huang, Xinghao Ding, Xiaoqing Liu, and Yizhou Yu. 2021. Generator versus segmentor: Pseudo-healthy synthesis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*. Springer, 150–160.

[86] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, YongJae Lee, Madison Madison, Microsoft Research, Redmond Hkust, Microsoft Cloud, and Ai Ai. [n. d.]. Segment Everything Everywhere All at Once. ([n. d.]).