UP-SAM: Uncertainty-Informed Adaptation of Segment Anything Model for Semi-Supervised Medical Image Segmentation

Wenjing Lu, Yi Hong*, and Yang Yang*
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
*Co-corresponding authors, Email: yi.hong@sjtu.edu.cn, yangyang@cs.sjtu.edu.cn

***Abstract*—Semi-supervised segmentation is extensively employed in medical image analysis due to its ability to leverage a small amount of labeled data alongside abundant unlabeled data. However, its performance is hindered by the inadequate knowledge of the data domain learned from limited labeled data and the absence of effective strategies for exploiting unlabeled regions, especially when annotations are extremely scarce. To address these challenges, the Segment Anything Model (SAM) has emerged as a promising solution. As a foundation model enriched by extensive and diverse domain knowledge, SAM has been leveraged to mitigate the epistemic uncertainty (EU) of semi-supervised segmentation models, while aleatoric uncertainty (AU) is often ignored. In this paper, we propose a novel semi-supervised medical image segmentation framework called UP-SAM, which adapts SAM for dual uncertainty assessments. The framework achieves effective collaboration between large foundation models and domain-specific models, leading to a simultaneous reduction in the impact of EU and AU. The experiments on the left atrium and pancreas datasets demonstrate the superior efficacy of UP-SAM against baseline methods. Particularly, UP-SAM exhibits substantial advantages over other semi-supervised learning models when dealing with exceedingly scarce labeled data. Code is available at https://github.com/VivienLu/UP-SAM.**

***Index Terms*—Segment Anything Model, Dual Uncertainty Assessments, Semi-Supervised Learning, Medical Image Segmentation**

I. Introduction

Medical image segmentation is crucial for accurate diagnostics and effective treatment planning, as it enables precise delineation of anatomical structures [1]. However, its impact is often limited by the scarcity of annotated data, which is costly and time-intensive to acquire. Semi-supervised learning (SSL) offers a promising solution by leveraging both labeled and unlabeled data, reducing annotation demands while enhancing generalization on complex medical images. Key SSL techniques include uncertainty estimation [2], pseudo-labeling [3], consistency regularization [4], and self-training [5].

Despite these advances, SSL for medical image segmentation faces persistent challenges, especially in managing uncertainties and effectively utilizing unlabeled data. Uncertainty can be divided into epistemic uncertainty (EU), caused by model limitations, and aleatoric uncertainty (AU), stemming from inherent data variability [6]. In SSL models, EU is often estimated to improve model reliability, using techniques like Monte Carlo dropout for uncertainty mapping [7], [8] and consistency regularization in teacher-student frameworks [9]. However, conventional methods tend to evaluate uncertainties on a per-pixel basis, assuming pixel independence. This approach overlooks the structured nature of AU in medical images, where inter-pixel label dependencies are critical [10]. As a result, while SSL models effectively capture EU, they often struggle to adequately represent AU, limiting their accuracy in complex medical image contexts.

To address AU estimation, recent approaches employ probabilistic models to capture spatial correlations, improving uncertainty estimates in fully supervised settings [10]–[12]. For SSL, models like FUSSNet [13] demonstrate that combining EU and AU enhances performance by applying pseudo-labels in reliable regions and consistency constraints on ambiguous ones. However, with limited labeled data, heavy reliance on unlabeled data can lead to inaccuracies, weakening the reliability of pseudo-labels [14]–[16].

The Segment Anything Model (SAM) [17] has made significant strides in medical segmentation [18], showcasing an exceptional ability to derive robust and generalizable features from images. To extend the capabilities of SAM in medical image analysis, MedSAM [18] trained the refined architecture from scratch on an extensive library of annotated medical masks, facilitating 3D segmentation via sequential 2D slice processing. A notable advancement, SAM-Med3D [19], enhances zero-shot 3D segmentation by using 3D positional encodings, achieving remarkable outcomes in intricate volumetric segmentation.

Despite these advances, SAM-based models often rely on manual prompts and struggle with fully unlabeled samples [18]. Integrations with domain-specific adaptations enhance performance in fields where medical expertise is limited. However, they frequently neglect AU due to dataset variability and cross-domain discrepancies [20], [21]. While recent approaches that combine EU and AU can improve segmentation outcomes, robust AU modeling within SAM remains underexplored, leaving a critical gap in effectively handling label variability across medical imaging datasets.

In this paper, we introduce a novel semi-supervised framework UP-SAM for medical image segmentation that exploits the full potential of SAM in addressing both EU and AU. ***UP-SAM aims to establish a harmonious collaboration***

between two distinct annotators: one with broad, generalized knowledge akin to the foundation model, and the other with focused, specialized expertise akin to the domain-specific model. Our main contributions are summarized below:

- We propose an uncertainty-informed semi-supervised framework that effectively adapt the SAM model to handle both epistemic and aleatoric uncertainties in medical image segmentation.
- To address EU, we combine diverse loss metrics and refine SAM-Med3D [19] using a minimal set of labeled samples, ensuring domain-specific adaptation while preserving the extensive foundational knowledge.
- To address AU, we employ stochastic modeling to align the logit distributions of the domain-specific model with SAM-Med3D masks, marking UP-SAM as the pioneering work in successfully implementing unsupervised AU estimation in the realm of SSL.
- Comprehensive experiments validate the efficacy of UP-SAM in semi-supervised CT and MRI segmentation, achieving superior performance even with extremely limited annotations.

II. Related Work

A. Semi-supervised Learning in Medical Image Segmentation

Semi-supervised learning (SSL) has advanced medical image segmentation by leveraging both labeled and unlabeled data. Techniques like consistency regularization, exemplified by the Mean Teacher (MT) model [22], stabilize predictions through temporal consistency. MCF [14] builds on MT by improving boundary precision through dynamic network interactions. Pseudo-labeling methods, such as MC-Net+ [8], enhance label quality using multi-decoder architectures for robust predictions, while self-training approaches like AC-MT [15] focus learning on ambiguous regions to capture the most informative data. However, these models face persistent challenges in managing uncertainty within unlabeled regions, balancing model complexity, and mitigating error propagation in pseudo-labeling. Addressing these limitations remains critical to enhancing the robustness and generalizability of SSL in clinical imaging applications.

B. Segmentation Foundation Models in Medical Imaging

The Segment Anything Model (SAM) [17], trained on extensive annotated datasets, has demonstrated strong potential in medical segmentation [18]. Its ability to extract generalizable features from medical images is further enhanced by innovations like 3D adapters [23], [24] and the generation of 3D prompts from 2D data [25]. SAM-Med3D [19] extends SAM into a fully 3D framework, leveraging over 131K 3D masks without fine-tuning SAM's pre-trained parameters, achieving impressive volumetric segmentation with minimal prompts.

Despite its strengths, SAM-based models have limitations when applied to fully unlabeled datasets, often depending on manually provided prompts to guide segmentation [18]. To address this, recent efforts integrate SAM with domain-specific models for improved segmentation. SAM-LST [26] combines SAM with CNNs, while SAMIHS [27] introduces a parameter-efficient method for intracranial hemorrhage segmentation. SemiSAM [20] and ASLseg [21] adapt SAM for SSL by incorporating domain knowledge and pseudo-labels, effectively reducing EU. However, challenges remain in addressing AU, which arises from discrepancies between general and domain-specific data.

C. Uncertainty Estimation in Medical Image Segmentation

Uncertainty estimation is essential for reliable semi-supervised medical image segmentation. Approaches like UA-MT [9] modulate consistency loss based on uncertainty between teacher and student models, enhancing adaptability to unlabeled data while risking error propagation from inaccurate pseudo-labels. URPC [16] mitigates this through a pyramid structure for uncertainty rectification, whereas UPCoL [28] utilizes entropy-based masks to reinforce prototype consistency across labeled and unlabeled data.

Nonetheless, many SSL methods assess pixel-wise uncertainty independently, overlooking structured uncertainty inherent in medical images, which is AU arising from inter-pixel dependencies. To address AU in fully supervised settings, methods such as U-Net with variational autoencoders [11], Bayesian deep learning [29], and stochastic segmentation networks [10] have been proposed. Zepf et al. [12] explore AU modeling based on labeling styles, while FUSSNet [13] incorporates both AU and EU in SSL, although it encounters limitations when multiple decoders produce consistently incorrect outputs.

To overcome these limitations, our UP-SAM directly targets EU by adapting SAM-Med3D to reduce errors from incorrect predictions. Additionally, it refines AU distribution in unlabeled data, improving the extraction and transfer of generalized knowledge.

III. Methodology

Let $\mathcal{D}$ denote the dataset, comprising a labeled subset $\mathcal{D}^L$ with N pairs (x_i^l, y_i^l), and an unlabeled subset $\mathcal{D}^U$ containing M instances x_i^u. Both subsets exist within a 3D space $\mathbb{R}^{H\times W\times D}$ and are segmented into C classes. The SAM-Med3D [19] model (p_s) and the domain-specific model (p_ϕ) predict segmentation masks $\hat{y}_s$ and $\hat{y}_\phi$, respectively. For labeled data, SAM-Med3D iteratively selects prompt points where $\hat{y}_s^l$ diverges from the ground truth y^l. For unlabeled data, $\hat{y}_s^u$ is derived from prompts where SAM-Med3D masks $\hat{y}_s^u$ differ from domain-specific model masks $\hat{y}_\phi^u$ in the prior iteration.

UP-SAM operates in two phases: pre-training and fine-tuning. Initially, the domain-specific model is pre-trained on a small number of labeled samples. During fine-tuning, UP-SAM incorporates two key components: 1) **Supervised Domain Adaptation**, which jointly fine-tunes SAM-Med3D and the domain-specific model, and 2) **Unsupervised Stochastic Alignment**, a stochastic modeling process aligning logit distributions from SAM-Med3D and the domain-specific model. The architecture is depicted in Fig. 1.

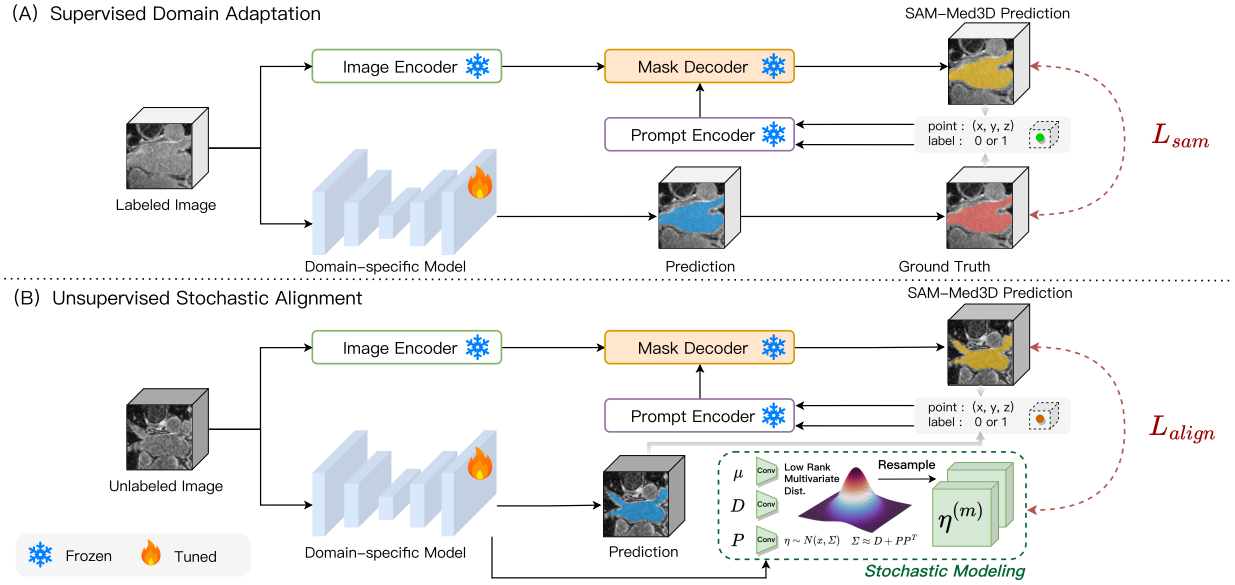

Fig. 1. Overview of our framework UP-SAM. (A) **Supervised Domain Adaptation**: SAM-Med3D is fine-tuned on labeled images, with segmentation supervised against ground truth, and further refined for domain-specific adaptation. (B) **Unsupervised Stochastic Alignment**: The logit distributions of the domain-specific model are aligned with SAM-Med3D's predictions using stochastic modeling on unlabeled data. The image encoder of SAM-Med3D is frozen, the mask decoder is partially tuned, while the domain-specific model is fully tuned.

A. Supervised Domain Adaptation (SDA)

Domain-specific Model Branch. The domain-specific model, compatible with architectures such as VNet [30], integrates multiple classification heads to capture diverse features and reduce EU. Its backbone includes a shared encoder and decoder, along with four EU heads constrained by various loss functions (Cross-Entropy, Dice, IoU, and Focal Loss [31]) to optimize classification accuracy, segmentation overlap, and class balance. Averaging predictions from these heads smooths outputs, mitigating overfitting and enhancing robustness.

During training, we apply a higher learning rate during pre-training to improve generalization, followed by a lower rate in fine-tuning for precision. The segmentation loss used in both phases is defined as:

$$L_{seg} = \sum_{h \in H} L_h(p_\phi^h(x_i^l), y_i^l). \tag{1}$$

where $H = \{ce, dice, focal, iou\}$ denotes the loss functions applied to each EU head.

SAM-Med3D Branch. We leverage SAM-Med3D [19] as the foundation model and fine-tune it on limited labeled samples using a combined Cross-Entropy and Dice loss function, following SAM-Med3D's settings. To maintain general knowledge and reduce computational costs, only the final layers of the mask decoder are updated (upscaling block and output MLP), while other layers remain frozen [26]. The fine-tuning objective is expressed as:

$$L_{sam} = L_{ce+dice}(p_s(x_i^l), y_i^l). \tag{2}$$

B. Unsupervised Stochastic Alignment (USA)

To address AU in unlabeled data, we align domain-specific model logits with SAM-Med3D's predictions for unlabeled images. Unlike previous approaches that depend on labeled data [10], [12], our unsupervised alignment module leverages unlabeled data to enhance generalization.

The stochastic modeling process generates multiple segmentation hypotheses for each unlabeled image, reducing AU by exploring plausible segmentations. Initial segmentations from SAM-Med3D undergo stochastic perturbations to produce diverse yet coherent hypotheses and capture inherent variability in medical images.

Following [10], we model the domain-specific logits η_ϕ as a multivariate normal distribution $N(\mu(x), \Sigma(x))$. To enhance computational efficiency, we represent $\Sigma(x)$ in a low-rank form $\Sigma(x) = D(x) + P(x)P(x)^T$, where $D(x)$ is a diagonal variance matrix and $P(x)$ provides low-rank covariance factors. Here, $\mu(x)$, $P(x)$, and $D(x)$ are outputs from distinct convolutional neural networks. Estimating the conditional log-likelihood involves K Monte Carlo samples to refine $\mu(x)$, $P(x)$, and $D(x)$, as follows:

$$\log p(y|x) = \log \int p(y|\eta_\phi) p_\theta(\eta_\phi|x) d\eta \approx \log \frac{1}{K} \sum_{k=1}^{K} p(y|\eta_\phi^{(k)}), \tag{3}$$

where $p_\theta(\eta_\phi|x)$ represents the logit probability distribution conditioned on the image, parameterized by θ.

Unlike prior work [10], which relies on ground truth, we utilize the SAM-Med3D's masks $\hat{y}_s^u$ for unlabeled images to calibrate the alignment loss, formulated as:

$$\mathcal{L}_{align} = -\sum_{k}^{K} \sum_{i}^{H \times W \times D} \sum_{c}^{C} \hat{y}_{s,ic}^u \log(\text{softmax}(\eta_{\phi,ic}^{(k)})). \tag{4}$$

We hypothesize that fine-tuning expands the intersection

between domain-specific and foundational knowledge, transitioning from generalized to specialized insights. The alignment ensures that the domain-specific model effectively integrates and leverages knowledge from SAM-Med3D, promoting consistent and robust segmentation.

Finally, the total loss of our UP-SAM framework, accounting for labeled and unlabeled data, is outlined as:

$$Loss = \alpha(L_{seg} + L_{sam}) + \lambda L_{align}, \tag{5}$$

where α is set to 0.5 to balance the labeled terms, and λ is a temporal Gaussian warming-up function to modulate alignment strength.

IV. Experiments

A. Datasets and Experimental Setup

Datasets. We evaluate the proposed framework on two widely used semi-supervised medical image segmentation datasets: the Left Atrium (LA) dataset [32] and the Pancreas-CT dataset [33]. The LA dataset contains 100 high-resolution 3D MR images, split into 80 for training and 20 for testing [7], [9], [34]–[36]. The Pancreas-CT dataset comprises 82 CT scans, with 62 for training and 20 for testing [7], [13], [28], [34]. Both datasets are normalized, with labeled subsets including 1, 2, and 4 scans, and the remainder used as unlabeled data following standard selection protocols [7], [34]. Data preprocessing contains extracting $128 \times 128 \times 128$ voxel cubes and applying one-hot encoding for multiclass masks, consistent with SAM-Med3D [19].

Implementation details. We use V-Net [30] as the backbone. The model undergoes pre-training for 7.5k iterations using SGD ($lr = 10^{-2}$) and fine-tuning for 7.5k iterations with AdamW ($lr = 10^{-5}$), following SSL standards [7]–[9], [14], [34] and SAM-Med3D's configuration [19]. Batch sizes are 4 (2 labeled, 2 unlabeled) or 2 (1 labeled, 1 unlabeled), depending on labeled scan availability. For SAM-Med3D inference, we use five prompt points per iteration, based on ground truth for labeled data or model predictions for unlabeled data. To model AU, we use 20 Monte Carlo samples [10], [13], and the alignment term λ follows a Gaussian ramp-up schedule with a maximum value of 1 [7], [16], [34].

Performance is evaluated with Dice, Jaccard Index, 95% Hausdorff Distance, and Average Symmetric Surface Distance, using PyTorch on an NVIDIA RTX 3090 GPU.

Baseline approaches. We compare our framework against both pre-trained foundation models and SSL methods using the V-Net [30] backbone for medical image segmentation. Baseline SSL methods include MT [22] and UA-MT [9] for consistency regularization and uncertainty estimation, MC-Net+ [8] for multi-decoder pseudo-labeling, FUSSNet [13] for integrating AU and EU, MCF [14] for dual-network interactions, and AC-MT [15] for selective consistency in ambiguous regions.

For zero-shot segmentation, we benchmark against SAM [17] and SAM-Med3D [19]. SAM requires manual bounding boxes per slice, while SAM-Med3D uses five random foreground prompts per volume. For fully supervised segmentation, MedSAM [18], pre-trained on datasets such as LA and Pancreas-CT, is applied.

TABLE I
Segmentation performance comparison on the Left Atrium dataset with varying labeled (#Lb) and unlabeled (#Unlb) data ratios.

Method	#Lb/#Unlb	Dice(%)	Jac(%)	95HD	ASD
Pretrained foundation models for Zero-shot Segmentation					
SAM	0 / 80	77.38	63.43	12.72	4.75
SAM-Med3D		77.96	64.78	13.03	3.73
Semi-Supervised Segmentation with Extremely Limited Labeled Data					
VNet	1 / 79	7.20	5.69	44.14	26.92
MT (NIPS'17)		44.16	29.66	47.02	**2.20**
UA-MT (MICCAI'19)		46.51	31.84	44.11	5.42
MC-Net+ (MIA'22)		4.09	2.92	37.26	9.98
FUSSNet (MICCAI'22)		65.88	50.23	40.69	13.92
MCF (CVPR'23)		5.33	4.42	32.41	19.88
AC-MT (MIA'23)		46.74	32.21	43.99	18.44
UP-SAM		**78.25**	**65.31**	**18.83**	4.39
VNet	2 / 78	46.74	32.60	39.02	6.24
MT (NIPS'17)		69.32	53.87	37.21	2.92
UA-MT (MICCAI'19)		73.38	58.82	31.94	2.69
MC-Net+ (MIA'22)		27.17	18.51	34.45	11.70
FUSSNet (MICCAI'22)		74.92	61.58	27.17	8.81
MCF (CVPR'23)		67.57	52.49	32.33	**2.60**
AC-MT (MIA'23)		81.33	68.87	16.50	5.16
UP-SAM		**82.84**	**71.25**	**16.03**	4.36
VNet	4 / 76	67.34	55.26	25.70	7.23
MT (NIPS'17)		74.64	60.77	34.14	2.72
UA-MT (MICCAI'19)		75.82	62.09	28.36	3.27
MC-Net+ (MIA'22)		78.66	65.88	22.27	6.04
FUSSNet (MICCAI'22)		81.97	70.46	20.54	5.94
MCF (CVPR'23)		82.73	71.32	16.59	2.28
AC-MT (MIA'23)		**86.04**	**76.03**	**9.23**	**2.50**
UP-SAM		84.06	72.90	13.78	3.14
Fully Supervised Segmentation					
VNet	80 / 0	91.42	84.27	5.15	1.50
MedSAM		81.34	68.73	10.47	3.71

B. Experimental Results on the Left Atrium Dataset

Table I demonstrates the effectiveness of our UP-SAM method on the LA dataset, particularly in low-annotation settings. With just 1 labeled scan, UP-SAM achieves a Dice score of 78.25%, outperforming other SSL methods. With 4 labeled scans, UP-SAM achieves a Dice score of 84.06%, slightly trailing AC-MT [15] at 86.04%. This may be due to the relative simplicity of the LA dataset, where domain-specific knowledge suffices, and excess general knowledge can introduce redundancy.

In zero-shot segmentation, SAM-Med3D [19] slightly outperforms SAM [17], benefiting from 3D positional encodings. Fine-tuning UP-SAM with 2 labeled scans yields a Dice score of 82.84%, surpassing both SAM (77.38%) and SAM-Med3D (77.96%). This highlights the advantage of UP-SAM's domain-specific adaptation and stochastic modeling.

In fully supervised settings, VNet achieves a Dice score of 91.42%, exceeding MedSAM's 81.34%. This result, consistent with prior studies [37], [38], underscores the importance of domain-specific adaptation for medical segmentation tasks.

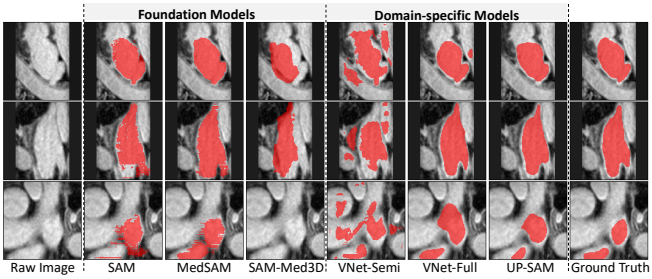

Fig. 2. Visual comparison of left atrium segmentation results (in red).

Figure 2 shows segmentation results using 2 labeled scans. While SAM and SAM-Med3D produce coarse segmentations with missing portions of the left atrium, MedSAM improves consistency yet struggles with boundary precision. VNet-Semi, trained on 2 labeled scans, displays significant errors, whereas VNet-Full, trained on the full dataset, reduces errors but still misidentifies some regions. In contrast, our UP-SAM provides accurate and detailed segmentations, closely aligning with ground truth even with limited labeled data.

C. Experimental Results on the Pancreas-CT Dataset

Tab. II presents a comparison of segmentation methods on the Pancreas-CT dataset, demonstrating the strong performance of UP-SAM. With just one labeled scan, UP-SAM achieves a Dice score of 70.79%, significantly outperforming FUSSNet's 30.67%. Although both UP-SAM and FUSSNet address EU and AU, UP-SAM's superior results highlight the added benefit of incorporating general knowledge from foundation models.

Moreover, UP-SAM's performance with a single labeled scan nearly matches MedSAM's fully supervised score with 62 labeled scans (70.92%). This demonstrates UP-SAM's efficiency in leveraging limited annotations and offering a cost-effective solution for medical image segmentation.

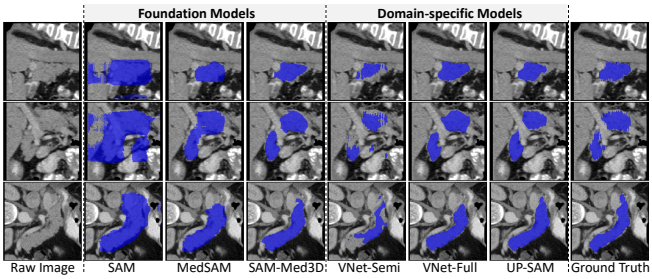

Fig. 3. Visual comparison of pancreas segmentation results (in blue).

Figure 3 provides a visual comparison of pancreas-CT segmentation results across different methods. The Pancreas-CT dataset presents challenges due to jagged and concave shapes in the annotations, complicating edge segmentation. Despite these challenges, UP-SAM captures the pancreas structure with high precision, delivering smooth and consistent boundaries. It closely matches the performance of fully supervised models like MedSAM and VNet-Full, even with limited labeled data, demonstrating its effectiveness in complex medical image segmentation tasks.

TABLE II
SEGMENTATION PERFORMANCE COMPARISON ON THE PANCREAS-CT DATASET WITH VARYING LABELED (#LB) AND UNLABELED (#UNLB) DATA RATIOS.

Method	#Lb/#Unlb	Dice(%)	Jac(%)	95HD	ASD
Pretrained foundation models for Zero-shot Segmentation					
SAM*	0 / 62	35.95	22.06	37.41	17.04
SAM-Med3D*		79.60	66.36	5.34	1.49
Semi-Supervised Segmentation with Extremely Limited Labeled Data					
VNet	1 / 61	19.94	11.86	43.29	11.38
MT (NIPS'17)		19.22	10.84	67.01	2.38
UA-MT (MICCAI'19)		9.51	5.03	70.06	2.69
MC-Net+ (MIA'22)		10.42	5.54	65.35	33.52
FUSSNet (MICCAI'22)		30.67	18.61	43.86	19.86
MCF (CVPR'23)		11.45	6.10	67.45	**2.15**
AC-MT (MIA'23)		15.94	9.18	56.85	30.23
UP-SAM		**70.79**	**55.25**	**11.00**	3.41
VNet	2 / 60	37.20	24.73	41.18	5.55
MT (NIPS'17)		37.34	23.66	51.17	2.49
UA-MT (MICCAI'19)		21.44	12.33	61.12	3.63
MC-Net+ (MIA'22)		17.57	9.82	59.81	29.32
FUSSNet (MICCAI'22)		40.53	26.10	45.95	18.86
MCF (CVPR'23)		19.35	10.85	63.64	**1.71**
AC-MT (MIA'23)		21.63	13.04	65.81	32.84
UP-SAM		**71.12**	**55.61**	**11.39**	3.38
VNet	4 / 58	50.01	34.46	41.01	3.67
MT (NIPS'17)		35.67	22.71	57.53	3.51
UA-MT (MICCAI'19)		34.60	22.31	53.38	3.63
MC-Net+ (MIA'22)		33.78	21.11	53.33	25.13
FUSSNet (MICCAI'22)		58.81	43.37	34.44	12.51
MCF (CVPR'23)		42.28	28.70	53.56	27.53
AC-MT (MIA'23)		46.42	31.36	51.71	2.94
UP-SAM		**72.65**	**57.42**	**10.49**	**2.82**
Fully Supervised Segmentation					
VNet	62 / 0	82.53	70.63	6.01	1.88
MedSAM*		70.92	55.09	7.71	2.68

TABLE III
ABLATION STUDY ON THE LEFT ATRIUM DATASET BY UTILIZING TWO LABELED SCANS.

Method		Backbone		Loss			Metrics	
		VNet	SAM-Med3D	L_{seg}	L_{sam}	L_{align}	Dice(%)	ASD
Sup-Only	Baseline	✓					46.74	6.24
	+ EU Heads	✓		✓			72.81	10.86
	+ SDA	✓	✓	✓	✓		77.78	7.10
Semi-Sup	+ USA	✓	✓	✓		✓	81.31	3.93
	UP-SAM	✓	✓	✓	✓	✓	82.84	4.36

D. Ablation Study

The ablation studies of key components in UP-SAM are summarized in Tab. III. In the supervised-only segmentation study, applying EU estimation to VNet (+ EU Heads) raised the Dice score from 46.74% to 72.81%, illustrating the value of leveraging EU heads for richer exploitation of labeled data. Further improvement was seen by introducing the Supervised Domain Adaptation module (+ SDA), raising the Dice score to 77.78%, demonstrating the importance of fine-tuning even well-established models. In the semi-supervised setting, adding the Unsupervised Stochastic Alignment module (+ USA) to model AU significantly boosted performance to 81.31%. This highlights the importance of modeling AU in

leveraging unlabeled data by applying generalized knowledge from the foundation model to the specific domain. Finally, the combination of supervised fine-tuning (L_{sam}) and unsupervised alignment (L_{align}) further improved the Dice score to 82.84%, confirming the complementary benefits of domain adaptation and uncertainty modeling for better leveraging both labeled and unlabeled data.

V. Conclusion

In this work, we propose a refined semi-supervised framework UP-SAM that seamlessly integrates epistemic and aleatoric uncertainty assesment within a dual-model architecture, specifically designed for medical image segmentation with scarce annotations. The framework encompasses a comprehensive strategy that involves fine-tuning SAM-Med3D to address EU and logit space alignment to handle AU, aiming to achieve collaboration between large foundation models and domain-specific models. Experimental results on two datasets demonstrate that UP-SAM outperforms most existing SSL and pre-trained foundation models, achieving high performance with minimal labeled scans and no need for prompts during inference. Future work will explore extending this framework to multi-class segmentation tasks and further refining it to minimize redundant knowledge interference.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC 62272300 and 62203303).

References

[1] S. Suganyadevi *et al.*, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 19–38, 2022.

[2] Y. Wang *et al.*, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4248–4257.

[3] Y. Shi *et al.*, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE transactions on medical imaging*, vol. 41, no. 3, pp. 608–620, 2021.

[4] W. Hang *et al.*, "Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation," in *MICCAI 2020*. Springer, 2020, pp. 562–571.

[5] X. Li *et al.*, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2020.

[6] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2021.

[7] Y. Wu *et al.*, "Semi-supervised left atrium segmentation with mutual consistency training," in *MICCAI 2021*. Springer, 2021, pp. 297–306.

[8] ——, "Mutual consistency learning for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 81, p. 102530, 2022.

[9] L. Yu *et al.*, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *MICCAI 2019*. Springer, 2019, pp. 605–613.

[10] M. Monteiro *et al.*, "Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty," *Advances in neural information processing systems*, vol. 33, pp. 12 756–12 767, 2020.

[11] S. Kohl *et al.*, "A probabilistic u-net for segmentation of ambiguous images," *Advances in neural information processing systems*, vol. 31, 2018.

[12] K. Zepf *et al.*, "That label's got style: Handling label style bias for uncertain image segmentation," *arXiv preprint arXiv:2303.15850*, 2023.

[13] J. Xiang *et al.*, "Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation," in *MICCAI 2022*. Springer, 2022, pp. 481–491.

[14] Y. Wang *et al.*, "Mcf: Mutual correction framework for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 651–15 660.

[15] Z. Xu *et al.*, "Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 88, p. 102880, 2023.

[16] X. Luo *et al.*, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, p. 102517, 2022.

[17] A. Kirillov *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[18] J. Ma *et al.*, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[19] H. Wang *et al.*, "Sam-med3d," *arXiv preprint arXiv:2310.15161*, 2023.

[20] Y. Zhang *et al.*, "Semisam: Exploring sam for enhancing semi-supervised medical image segmentation with extremely limited annotations," *arXiv preprint arXiv:2312.06316*, 2023.

[21] S. Chen *et al.*, "Aslseg: Adapting sam in the loop for semi-supervised liver tumor segmentation," *arXiv preprint arXiv:2312.07969*, 2023.

[22] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[23] J. Wu *et al.*, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[24] C. Chen *et al.*, "Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation," *arXiv preprint arXiv:2309.08842*, 2023.

[25] C. Wang *et al.*, "Sam med: A medical image annotation framework based on large vision model," *arXiv preprint arXiv:2307.05617*, vol. 3, 2023.

[26] S. Chai *et al.*, "Ladder fine-tuning approach for sam integrating complementary network," *arXiv preprint arXiv:2306.12737*, 2023.

[27] Y. Wang *et al.*, "Samihs: Adaptation of segment anything model for intracranial hemorrhage segmentation," *arXiv preprint arXiv:2311.08190*, 2023.

[28] W. Lu *et al.*, "Upcol: Uncertainty-informed prototype consistency learning for semi-supervised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 662–672.

[29] A. Kendall *et al.*, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[30] F. Milletari *et al.*, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[31] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[32] Z. Xiong *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical image analysis*, vol. 67, p. 101832, 2021.

[33] H. R. Roth *et al.*, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *MICCAI 2015*. Springer, 2015, pp. 556–564.

[34] S. Li *et al.*, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *MICCAI 2020*. Springer, 2020, pp. 552–561.

[35] C. You *et al.*, "Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2228–2237, 2022.

[36] T. Lei *et al.*, "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *IEEE Transactions on Medical Imaging*, 2022.

[37] B. Glocker *et al.*, "Risk of bias in chest radiography deep learning foundation models," *Radiology: Artificial Intelligence*, vol. 5, no. 6, p. e230060, 2023.

[38] Y. Huang *et al.*, "Segment anything model for medical images?" *Medical Image Analysis*, vol. 92, p. 103061, 2024.