Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2024 February ; 12926: . doi:10.1117/12.3006867.

FNPC-SAM: Uncertainty-Guided False Negative/Positive Control for SAM on Noisy Medical Images

Xing Yao[a], **Han Liu**[a], **Dewei Hu**[b], **Daiwei Lu**[a], **Ange Lou**[b], **Hao Li**[b], **Ruining Deng**[a], **Gabriel Arenas**[d], **Baris Oguz**[d], **Nadav Schwartz**[d], **Brett C Byram**[c], **Ipek Oguz**[a]

[a]Dept. of Computer Science, Vanderbilt University, Nashville, TN, USA

[b]Dept. of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, USA

[c]Dept. of Biomedical Engineering, Vanderbilt University, Nashville, TN, USA

[d]Dept. of Obstetrics and Gynecology, University of Pennsylvania, Philadelphia, PA, USA

Abstract

The Segment Anything Model (SAM) is a recently developed all-range foundation model for image segmentation. It can use sparse manual prompts such as bounding boxes to generate pixel-level segmentation in natural images but struggles in medical images such as low-contrast, noisy ultrasound images. We propose a refined test-phase prompt augmentation technique designed to improve SAM's performance in medical image segmentation. The method couples multi-box prompt augmentation and an aleatoric uncertainty-based false-negative (FN) and false-positive (FP) correction (FNPC) strategy. We evaluate the method on two ultrasound datasets and show improvement in SAM's performance and robustness to inaccurate prompts, without the necessity for further training or tuning. Moreover, we present the Single-Slice-to-Volume (SS2V) method, enabling 3D pixel-level segmentation using only the bounding box annotation from a single 2D slice. Our results allow efficient use of SAM in even noisy, low-contrast medical images. The source code has been released at: https://github.com/MedICL-VU/FNPC-SAM

Keywords

SAM; zero-shot; uncertainty; medical image segmentation; prompt engineering; kidney; placenta; ultrasound

1. INTRODUCTION

The performance of deep convolutional neural networks (DNNs) relies on large datasets with pixel-level annotations, which can be both time-consuming and labor-intensive to obtain. To address this, researchers have explored coarse-to-fine segmentation, learning pixel-level masks from easier-to-obtain annotations like bounding boxes (BBs).[1]

Recently, the Segment Anything Model (SAM)[2] proposed by Meta AI has gained significant attention as an all-range segmentation foundation model, capable of generating fine-grade segmentation masks using sparse prompts like BBs, fore/background points, texts, and dense prompts such as masks. While SAM excels in natural image segmentation,[3-5] its

performance in various medical image segmentation tasks can be unsatisfactory,[6-9] leading to increased interest in enhancing its capabilities in medical imaging scenarios.[10-16]

Efforts to improve SAM's performance on medical images fall into three categories: 1) Fine-tuning a modified SAM model with extensive medical image datasets.[10, 17] This requires additional time and effort for annotation and training, and the results depend on data availability and task complexity. 2) Directly using SAM's predicted results as fake ground truth to train a new network.[12] This approach relies on SAM's initial performance on the specific task. 3) Leveraging prompt optimization or augmentation strategies to achieve zero-shot segmentation.[11, 13, 15] These strategies can efficiently enhance zero-shot segmentation performance of SAM, with the potential to be leveraged across unseen datasets and downstream tasks.

Inspired by,[15, 18] we propose a test-phase prompt augmentation method to amplify the coarse-to-fine zero-shot segmentation performance of SAM on low-contrast and noisy ultrasound images, without supplementary training or fine-tuning. We assess the performance on kidney and placenta ultrasound images. Our results indicate that our proposed method surpasses the performance of standalone SAM and, notably, exhibits exceptional robustness to variations in the prompt.

We have 4 main contributions: **1)** SAM's performance is sensitive to the BB position and size,[6, 15, 19] and prediction based on a single prompt may contain FP and FN regions, as shown in Fig. 1(a). We suggest a Monte Carlo BB sampling approach to provide additional foreground/background prompts to SAM. The predictions from different field-of-view (FOV) allow us to estimate the aleatoric uncertainty map.[20] **2)** In previous studies, the uncertainty map is only used to represent the reliability and robustness of segmentation. Here, we further leverage the aleatoric uncertainty map for FN and FP correction (FNPC) in the averaged predictions. **3)** We examine the influence of the BB prompts in our study in fine (tight) BB, medium BB, and coarse BB scenarios. **4)** We introduce a Single-Slice-to-Volume (SS2V) approach for extension to 3D. This allows for pixel-level segmentation of an entire 3D volume based solely on the BB annotation of a single 2D slice.

2. METHODS

2.1 Monte Carlo bounding box sampling strategy

Combining SAM with bounding boxes as prompts has shown good segmentation results in medical images, although the performance is sensitive to the BB position and size.[6, 15, 19] SAM-U[15] addresses this by implementing a multi-BB prompt augmentation approach. While SAM-U only uses a simple random sampling strategy, we propose a Monte Carlo sampling method with constrained positions, as shown in Fig. 1(a). Our sampling method involves randomly selecting N points within a radius $R = \frac{1}{P}$ min(BB edges) around the center of the initial BB to generate new BBs of the same size. P is the radius ratio. This sampling strategy serves two purposes: first, the new sampled BBs (Fig. 1(a)) provide additional positive and negative prompts to SAM. Second, our sampling approach emulates the imprecision of typical manual BB placement.

The predictions M_i from augmented BBs are averaged as $M_{ave} = \{\frac{1}{N+1}\sum_{i=1}^{N+1} M_i\} > T_{ave}$. While SAM-U uses a threshold of $T_{ave} = 0$, effectively taking the union of the predictions, we use $T_{ave} = 0.5$ for a majority vote approach.

2.2 Uncertainty Estimation

We propose a straightforward approach to compute aleatoric uncertainty[20] UC_{raw} from the $N + 1$ predictions of augmented BBs and the initial BB. By analyzing the frequency f of foreground pixels in the set of predicted masks $\{M_i\}$, the uncertainty associated with each pixel in position (j, k) is determined by $f(j,k) = \frac{1}{N+1}\sum_{i=1}^{N+1} M_i(j,k)$. This frequency calculation allows us to compute the aleatoric uncertainty UC (we drop the pixel coordinates from the notation for brevity):

$$UC_{raw} = f \cdot (1 - f) \tag{1}$$

For a more entropy-like uncertainty, we propose the following calculation, with $\epsilon = 10^{-7}$:

$$UC_{raw} = -0.5 \cdot [f \cdot \log(f + \epsilon) + (1 - f) \cdot \log(1 - f + \epsilon)] \tag{2}$$

In our experiments, Equations 1 and 2 share the same performance. Finally, we threshold to extract the high-uncertainty areas: $UC = UC_{raw} > [min(UC_{raw}) + T_{UC} \cdot (max(UC_{raw}) - min(UC_{raw}))]$, where T_{UC} is the threshold ratio.

2.3 False Negative and False Positive Correction (FNPC)

Fig. 1(b) illustrates the pipeline of our proposed FNPC strategy. Given an input image I, we determine the average prediction M_{ave} (Sec. 2.1) and the uncertainty map UC which highlights potential FNs and FPs (Sec. 2.2).

False Negative Correction: Our goal is to identify FNs that are outside M_{ave} but within UC. Initially, the potential FN mask M_{PFN} is computed as $(1 - M_{ave}) \cdot UC$, and its corresponding regions in I are extracted as $I_{PFN} = I \cdot M_{PFN}$. To encourage intensity homogeneity, only pixels in I_{PFN} with intensities within a range $[T_{FNl}, T_{FNh}]$ are kept in the final FN mask M_{FN}: $M_{FN} = M_{PFN} \cdot (T_{FNh} > I_{PFN} > T_{FNl})$.

False Positive Correction: We aim to identify FPs present in both M_{ave} and UC. The potential FP mask M_{PFP} is derived from $M_{ave} \cdot UC$, with associated intensity values $I_{PFP} = I \cdot M_{PFP}$. The final FP mask M_{FP} is again determined with an intensity range $[T_{FPl}, T_{FPh}]$: $M_{FP} = M_{PFP} \cdot (T_{FPh} < I_{PFP} \| I_{PFP} < T_{FPl})$.

The final mask M_{Final} which corrects for FNs and FPs is given by: $M_{FNPC} = M_{ave} + M_{FN} - M_{FP}$. As shown in Fig. 1(a), the FNPC has good FN and FP correction performance compared to both the raw SAM and simple averaging.

2.4 Single Slice to Volume method (SS2V)

We next introduce the **SS2V** method for pixel-level segmentation of a 3D volume only using a single 2D BB annotation. The workflow of **SS2V** is depicted in Fig. 2.

1. We first select a 2D slice K containing the target. A manual BB, labelled as **Box K**, is provided. While in theory our method can start from any slice containing the target, we begin with the central slice, as this typically offers a more representative view of the anatomy than a far-off side.

2. Using **Box K** as the initial BB, the segmentation **Pred K** is derived using our FNPC method. Based on **Pred K**, we generate a tightly fitting BB, **CBox K**, to be used as the candidate BB for the next slice.

3. We refine the **CBox K** based on **Box K** to generate the BB (**Box K+1** or **Box K−1**) for the neighboring slices.

4. Steps 2 and 3 are iteratively applied to produce BBs for subsequent slices until the whole volume is segmented.

We assume that transitions between neighboring slices should be smooth. We enforce this by comparing **CBox K** and **Box K** and restricting the movement of each corner by a threshold T_B in Step 3. For example, for the lower-left corner of **Box K+1**, we use $xmin_{BK+1} = xmin_{CBK}$ if $| xmin_{CBK} - xmin_{BK} | \leq T_B$, and $xmin_{BK+1} = xmin_{BK}$ otherwise.

2.5 Datasets, preprocessing, and implementation details

Kidney: We use a dataset of free-hand kidney ultrasound images. It comprises 124 samples from 9 subjects, each 128x128 in dimension, with manual segmentations. The images are normalized to [0, 255] range. We compute **Fine BB** as the tightest BB of the pixel-level masks and randomly expanding the edges outwards by 0 to 2 pixels. **Medium BB** and **Coarse BB** are produced by similar expansions, ranging from 2 to 4 pixels and 4 to 6 pixels, respectively.

Placenta: We use a 3D placenta ultrasound dataset from 4 subjects, with manual annotations. The images are resized to 128x128x128 and normalized to [0, 255] range. We extract all 2D coronal slices containing placenta, for a total of 222 2D images. For brevity, we only show the **Fine BB** setting. For **SS2V**, we use the Fine BB of the central slice as initial BB.

Hyperparameter selection: We use the pretrained ViT-L SAM model[2] to obtain the initial SAM segmentation. For kidneys, we randomly pick 14 images from one subject for hyperparameter setting. T_{UC} is 0.9 for the Fine BB, 0.1 for the Medium BB and Coarse BB. For all stages, P is 8, N is 30, T_{FNl} and T_{FPl} are both 0, T_{FNh} and T_{FPh} are both 20. For the

placenta segmentation and SS2V task, we extract the central slice of each subject for tuning the hyperparameter for the entire dataset. For both tasks, T_{UC} is 0.2, P is 4, N is 30, T_{FNl} and T_{FPl} are both 70, T_{FNh} and T_{FPh} are both 200. For SS2V, the T_B is 2.

3. RESULTS

3.1 Kidney dataset

Left panels of Fig. 3 and Table 1 compare SAM, Average, and FNPC on kidney images under three levels of prompt coarseness, qualitatively and quantitatively. FNPC effectively eliminates the FP portions within the average predictions and SAM, and improves the Dice, ASSD, and HD results, across all BB coarseness levels. We observe that unlike SAM and average predictions, FNPC only shows a small deterioration in Dice and ASSD between fine and coarse prompts. This highlights FNPC's robustness to prompt coarseness.

3.2 Placenta dataset and SS2V experiments

The right panel of Fig. 3 presents the results for 3D placenta segmentation using the SS2V method. The top four rows depict 2D segmentation results, using a manual Fine BB annotation (yellow box). The FNPC method delivers overall superior segmentation results, with fewer FPs and FNs than the SAM and Average methods. Remarkably, SS2V showcases performance on par with FNPC across all slices, even though it only uses a single 2D Fine BB annotation (yellow box) for the entire 3D segmentation task. We observe that synthetic BBs (blue boxes) produced by SS2V are less precise than the manual Fine Box annotations, especially as we move further away from the reference slice 66. Nonetheless, even with these coarse annotations, SS2V manages to yield segmentations that surpass those produced by the SAM and Average models using a new manual BB annotations for each slice, and approaches FNPC performance. This behavior leverages FNPC's robustness to BB variations. The right panel of Table 1 provides quantitative results, illustrating that FNPC has better performance than SAM and Average. SS2V's outcomes are closely aligned with those of FNPC, showcasing excellent extension to 3D segmentation tasks.

4. DISCUSSION AND CONCLUSION

We introduced a test-phase prompt augmentation method to adapt SAM to challenging medical image segmentation tasks, specifically for ultrasound images marked by low contrast and noise. This method, leveraging multi-box prompt augmentation and aleatoric uncertainty thresholding, aims to mitigate SAM's FN and FP predictions without requiring time-consuming pixel-level annotations. Our evaluation on two ultrasound datasets showcases substantial improvements in SAM's performance and robustness to prompt coarseness. We recognize, however, that continued exploration and optimization are required for dealing with increasingly complex and variable data.

We further proposed SS2V to produce 3D segmentations from a single 2D BB input, with excellent results. We note that while the HD metric is slightly better in SAM for the placenta experiment, it also has a large standard deviation. In contrast, the Dice and ASSD metrics are substantially better for the FNPC and SS2V methods.

ACKNOWLEDGMENTS

This work is supported, in part, by NIH R01HD109739 and NIH R01HL156034. Here we also express our thanks for the computational resource support by Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University.

REFERENCES

[1]. Xu Y, Gong M, Xie S, and Batmanghelich K, “Box-adapt: Domain-adaptive medical image segmentation using bounding boxsupervision,” arXiv preprint arXiv:2108.08432 (2021).

[2]. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollar P, and Girshick R, “Segment anything,” in [Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)], 4015–4026 (October 2023).

[3]. Yang J, Gao M, Li Z, Gao S, Wang F, and Zheng F, “Track anything: Segment anything meets videos,” arXiv preprint arXiv:2304.11968 (2023).

[4]. Yu T, Feng R, Feng R, Liu J, Jin X, Zeng W, and Chen Z, “Inpaint anything: Segment anything meets image inpainting,” arXiv preprint arXiv:2304.06790 (2023).

[5]. Li F, Zhang H, Sun P, Zou X, Liu S, Yang J, Li C, Zhang L, and Gao J, “Semantic-sam: Segment and recognize anything at any granularity,” arXiv preprint arXiv:2307.04767 (2023).

[6]. Deng R, Cui C, Liu Q, Yao T, Remedios LW, Bao S, Landman BA, Tang Y, Wheless LE, Coburn LA, et al. , “Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging,” in [Medical Imaging with Deep Learning, short paper track], (2023).

[7]. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, and Zhang Y, “Segment anything model for medical image analysis: an experimental study,” Medical Image Analysis 89, 102918 (2023). [PubMed: 37595404]

[8]. Wald T, Roy S, Koehler G, Disch N, Rokuss MR, Holzschuh J, Zimmerer D, and Maier-Hein K, “Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model,” in [Medical Imaging with Deep Learning, short paper track], (2023).

[9]. Mattjie C, de Moura LV, Ravazio RC, Kupssinskü LS, Parraga O, Delucis MM, and Barros RC, “Exploring the zero-shot capabilities of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guideline,” arXiv preprint arXiv:2305.00109 (2023).

[10]. Ma J, He Y, Li F, Han L, You C, and Wang B, “Segment anything in medical images.,” Nat Commun 15, 654 (2024). [PubMed: 38253604]

[11]. Zhang Y, Zhou T, Wang S, Liang P, Zhang Y, and Chen DZ, “Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model,” in [Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops], 129–139, Springer Nature Switzerland, Cham (2023).

[12]. Cui C and Deng R, “All-in-sam: from weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning,” in [Asia Conference on Computers and Communications, ACCC], (2023).

[13]. Shen C, Li W, Zhang Y, Wang Y, and Wang X, “Temporally-extended prompts optimization for sam in interactive medical image segmentation,” in [2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)], 3550–3557, IEEE (2023).

[14]. Lou A, Li Y, Yao X, Zhang Y, and Noble J, “Samsnerf: Segment anything model (sam) guides dynamic surgical scene reconstruction by neural radiance field (nerf),” arXiv preprint arXiv:2308.11774 (2023).

[15]. Deng G, Zou K, Ren K, Wang M, Yuan X, Ying S, and Fu H, “Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image,” arXiv preprint arXiv:2307.04973 (2023).

[16]. Lee HH, Gu Y, Zhao T, Xu Y, Yang J, Usuyama N, Wong C, Wei M, Landman BA, Huo Y, et al. , “Foundation models for biomedical image segmentation: A survey,” arXiv preprint arXiv:2401.07654 (2024).

[17]. Wu J, Fu R, Fang H, Liu Y, Wang Z, Xu Y, Jin Y, and Arbel T, “Medical sam adapter: Adapting segment anything model for medical image segmentation,” arXiv preprint arXiv:2304.12620 (2023).

[18]. Shin H, Kim H, Kim S, Jun Y, Eo T, and Hwang D, “Sdc-uda: Volumetric unsupervised domain adaptation framework for slice-direction continuous cross-modality medical image segmentation,” in [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition], 7412–7421 (2023).

[19]. Cheng D, Qin Z, Jiang Z, Zhang S, Lao Q, and Li K, “Sam on medical images: A comprehensive study on three prompt modes,” arXiv preprint arXiv:2305.00035 (2023).

[20]. Kendall A and Gal Y, “What uncertainties do we need in bayesian deep learning for computer vision?,” Advances in neural information processing systems 30 (2017).

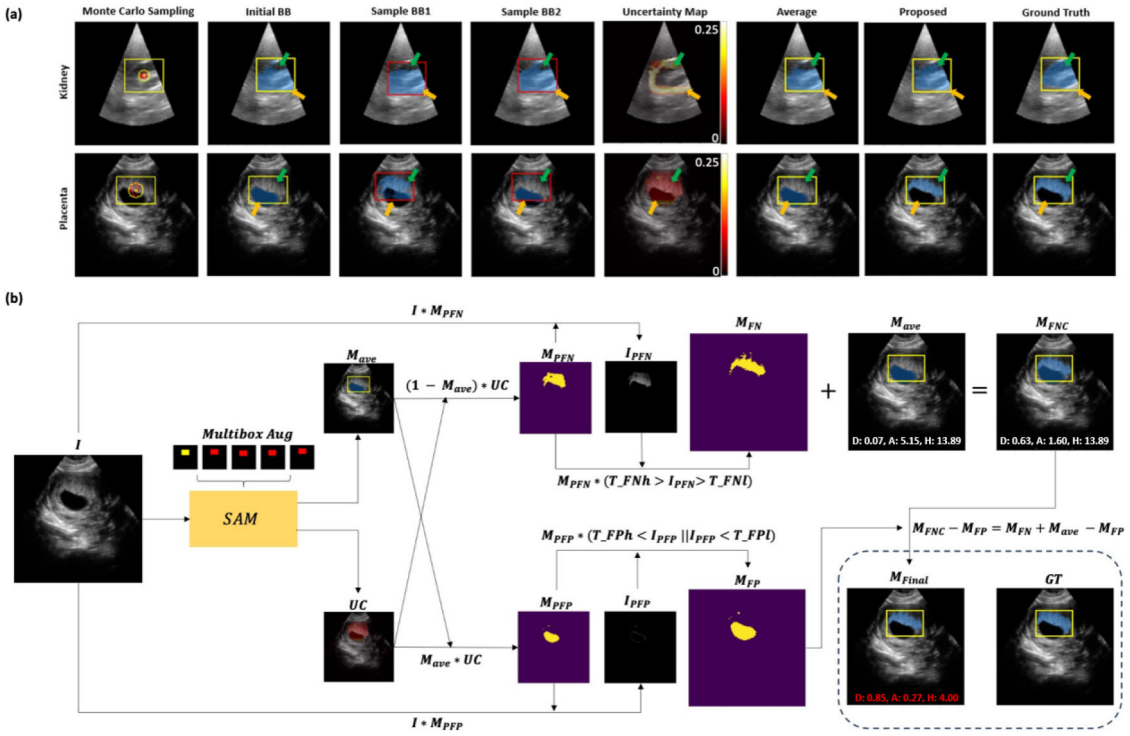

Figure 1.

(a) Monte Carlo BB sampling strategy (Sec. 2.1) on two segmentation tasks. The initial BB (yellow) has a center (yellow dot) and a sampling range (yellow circle). Sampled boxes (red) with their centers (red dots) are used to generate an average prediction. Green arrows point to FN areas in results from only the initial BB or simple averaging method, and are corrected by the proposed FNPC strategy. Orange arrows indicate areas where FPs are similarly corrected by the proposed FNPC. **(b)** FNPC pipeline (Sec. 2.3).

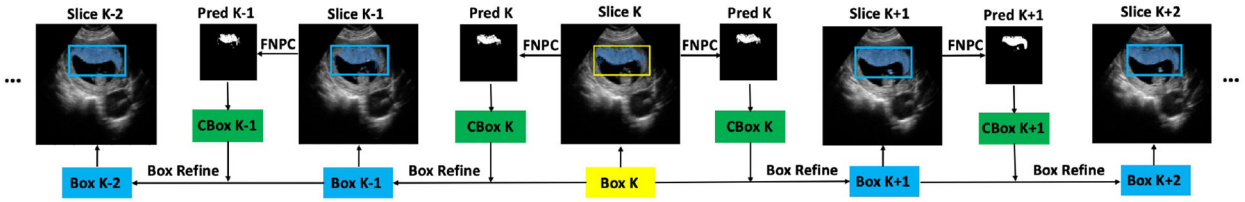

Figure 2.

Pipeline for **SS2V** method. Yellow color highlights the initial human-annotated BB. Green color highlights the candidate BBs generated from the predicted masks. Blue color highlights synthetic BBs generated by the **SS2V** method.

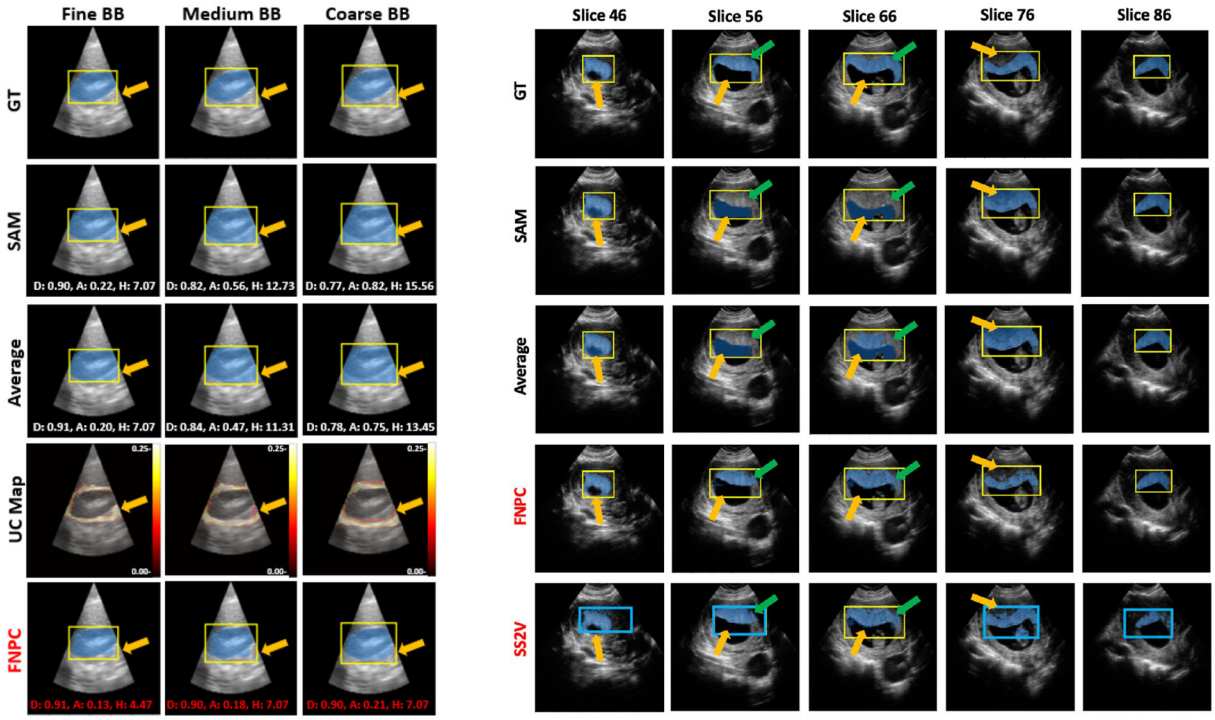

Figure 3.

Qualitative results. **Left,** kidney segmentation for three levels of prompt coarseness. D: Dice, A: ASSD, H: HD. Red font indicates the best performance. **Right,** placenta segmentation with SS2V. Yellow: manual BBs, blue: synthesized BBs generated by SS2V. For **Left** and **Right**, green and orange arrows highlight the FNs and FPs improved by our proposed methods (highlighted in red).

Table 1.

Quantitative results. F, M, C represent Fine BB, Medium BB, and Coarse BB, respectively. **Left,** kidney. Red numbers indicate the best performance for each BB scenario. **Right,** placenta with SS2V. Red numbers indicate the best performance for each metric.

Kidney	BB	Dice↑	ASSD↓	HD↓
SAM	F	0.90±0.05	0.25±0.21	6.67±3.44
	M	0.87±0.06	0.35±0.29	8.01±3.78
	C	0.77±0.06	0.83±0.37	12.96±4.12
Ave	F	0.90±0.05	0.23±0.20	6.26±3.52
	M	0.87±0.06	0.33±0.29	7.62±3.63
	C	0.77±0.05	0.77±0.34	12.27±3.89
FNPC	F	0.91±0.04	0.20±0.17	5.70±3.21
	M	0.89±0.05	0.24±0.22	5.87±3.09
	C	0.88±0.04	0.31±0.19	8.22±3.14

Placenta	BB	Dice↑	ASSD↓	HD↓
SAM	F	0.72 ± 0.13	0.84 ± 0.58	13.87 ± 1.99
Ave	F	0.70 ± 0.10	0.89 ± 0.28	14.96 ± 2.90
FNPC	F	0.79 ± 0.04	0.53 ± 0.18	14.66 ± 3.26
SS2V	F	0.76 ± 0.03	0.76 ± 0.19	15.55 ± 2.15