VISCERA-SAM: Adapting Segment Anything for Multi-Visceral Fetal Abdominal Ultrasound Segmentation

Minh H. N. Le[*1,2] D142111009@TMU.EDU.TW, JOHNMINHLE@IEEE.ORG
[1] *International Ph.D. Program in Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan*
[2] *AIBioMed Research Group, Taipei Medical University, Taipei, Taiwan*

Khoa D. Pham[*3] KDPHAM@AGGIES.NCAT.EDU
[3] *Industrial and Systems Engineering Department, North Carolina A&T State University, Greensboro, NC 27411, USA*

Tuan Vinh[4] TUAN.VINH@HERTFORD.OX.AC.UK
[4] *Medical Sciences Division, University of Oxford, Oxford, United Kingdom*

Thanh-Huy Nguyen[5] THANHHUN@ANDREW.CMU.EDU
[5] *Computational Biology Department, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213*

Han H. Huynh[6] M658112001@TMU.EDU.TW
[6] *International Master Program for Translational Science, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan*

Khanh T. Q. Le[7] KHANH.TRAN@UIT.EDU.VN
[7] *PASSIO Lab, North Carolina A&T State University, NorthGreensboro, NC 27411, USA*

Anh Mai Vu[8] MVU9@COUGARNET.UH.EDU
[8] *Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204, USA*

Dang Nguyen[9] KEVIN_NGUYEN@HSPH.HARVARD.EDU
[9] *Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA; Center for Materials Innovation and Technology, VinUniversity, Hanoi, Vietnam.*

Ha N. T. Luong[10] MVU9@COUGARNET.UH.EDU
[10] *Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204, USA*

Ulas Bagci[11] ULAS.BAGCI@NORTHWESTERN.EDU
[11] *Department of Radiology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.*

Min Xu[5] MXU1@ANDREW.CMU.EDU
[5] *Computational Biology Department, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213*

Carl Yang[12] J.CARLYANG@EMORY.EDU
[12] *Department of Computer Science, Emory University, Atlanta, GA, USA*

Phat K. Huynh[3] PKHUYNH@NCAT.EDU
[3] *Industrial and Systems Engineering Department, North Carolina A&T State University, Greensboro, NC 27411, USA*

Nguyen Quoc Khanh Le[13] KHANHLEE@TMU.EDU.TW
[13] *In-Service Master Program in Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, Taiwan; Translational Imaging Research Center, Taipei Medical University Hospital, Taipei, Taiwan; AIBioMed Research Group, Taipei Medical University, Taipei, Taiwan*

Editors: Under Review for MIDL 2026

Abstract

Foundation models such as the Segment Anything Model (SAM) have recently been explored for medical image segmentation, but their performance on fetal ultrasound, characterised by speckle noise, low contrast, and weak anatomical boundaries, remains insufficiently studied. We introduce VISCERA-SAM, a boundary-aware adaptation of the ultrasound-specific SAM-style foundation model UltraSAM for multi-organ fetal abdominal segmentation (abdominal aorta artery, intrahepatic umbilical vein, stomach, and liver area). Using UltraSAM architecture, fine-tuned to segment four clinically relevant fetal structures from single-point prompts. The approach enhances the baseline with (i) boundary-focused losses (Hausdorff or boundary loss) in addition to Dice-focal training, (ii) largest-connected-component filtering to suppress spurious regions, and (iii) geometry-preserving augmentations to improve robustness to venous shape variability. On our fetal abdominal dataset, the UltraSAM baseline achieves a mean Dice of 0.884 and mean Hausdorff distance of 12.1. VISCERA-SAM improves these results to a mean Dice of 0.910 while reducing mean Hausdorff distance by approximately 30% to 8.44. These findings indicate that ultrasound-native SAM-style pretraining, combined with lightweight boundary-aware adaptation, can provide accurate and contour-faithful fetal abdominal segmentations suitable for downstream biometric analysis.

Keywords: Fetal ultrasound, Abdominal organ segmentation, Foundation models, Segment Anything Model, Boundary-aware learning

1. Introduction

Fetal ultrasound (US) is central to prenatal care, enabling routine assessment of fetal growth and anatomy. A key biomarker is the Abdominal Circumference (AC), the most sensitive parameter for detecting fetal growth restriction and macrosomia (Salomon et al., 2019; Hadlock et al., 1985). Accurate AC measurement requires precise delineation of abdominal structures on a standardized transverse plane defined by the stomach, intrahepatic umbilical vein, and portal sinus (Salomon et al., 2019).

Manual AC measurement is operator dependent and highly sensitive to boundary quality, with errors amplified by speckle noise, attenuation, and variably defined contours. Automated segmentation methods must therefore maintain high boundary fidelity, since small contour deviations can lead to clinically significant differences in circumference.

Ultrasound segmentation remains challenging due to modality-specific artifacts, speckle noise, acoustic shadowing, attenuation, and boundary dropout, which reduce edge contrast and degrade performance relative to CT or MRI (Noble and Boukerroui, 2006). CNN-based approaches improve accuracy but typically require large task-specific datasets and lack generalization to unseen structures.

Foundation Models have introduced prompt-based segmentation at scale. The Segment Anything Model (SAM) (Kirillov et al., 2023) demonstrates impressive generalization on natural images, yet performs poorly on ultrasound because its training distribution differs fundamentally from sonographic textures. UltraSAM (Meyer et al., 2025) narrows this gap by pretraining on large ultrasound datasets, but still relies primarily on region-based losses

* Contributed equally

such as Dice, which do not sufficiently constrain boundary geometry. High-Dice predictions still contain jagged or irregular contours, limiting clinical utility (Kervadec et al., 2019).

We introduce **VISCERA-SAM**, a boundary-aware adaptation of UltraSAM designed for multi-organ fetal abdominal ultrasound segmentation. The method strengthens contour fidelity and improves robustness in noisy sonographic environments by integrating explicit boundary-oriented supervision and targeted artifact suppression. In particular, VISCERA-SAM incorporates both Boundary Loss and a differentiable Hausdorff Distance Loss to constrain contour geometry more precisely, addressing the weak and irregular edges characteristic of ultrasound. In addition, a Largest Connected Component (LCC) post-processing step is employed to eliminate speckle-induced false positives and isolated noise regions, ensuring cleaner and anatomically coherent predictions.

2. Methodology

2.1. Study Population

We used the publicly available UFSC *Fetal Abdominal Structures Segmentation Dataset* (Da et al., 2023), comprising nearly 1500 abdominal ultrasound images from 169 pregnant women collected between September 2021 and September 2023 at the University Hospital Polydoro Ernani de São Thiago, Florianópolis, Brazil.

Eligible participants were term pregnant women (age ≥ 18) undergoing labor, induction, or scheduled delivery, including cases with rupture of membranes, gestational diabetes, pre-eclampsia, or intrauterine growth restriction. Exclusion criteria included preterm gestations, multiple pregnancies, and known fetal structural or chromosomal anomalies. Gestational age was confirmed from the earliest ultrasound exam.

2.2. Data Preprocessing and Annotation Format

The UFSC fetal abdominal segmentation dataset (Da et al., 2023) provides DICOM ultrasound images together with NumPy (`.npy`) masks exported from 3D Slicer, where each file stores a dictionary containing pixel-wise annotations for the aorta, intrahepatic umbilical vein, stomach, and liver. All DICOMs were first converted to grayscale PNGs, rescaled to 8-bit intensity, and normalized to zero mean and unit variance. The structure-specific masks were then decoded and merged into a consistent multi-class label map. To satisfy SAM-style input requirements, both images and masks were padded and resized to a standardized resolution of 1024×1024 while preserving aspect ratio.

Following preprocessing, all masks were converted into COCO-style JSON annotations, with each organ represented by polygonal contours and assigned a unique category identifier. A subject-level 70/15/15 split was applied to form the training, validation, and testing sets, ensuring that images from the same patient did not appear across different partitions. Finally, quality-control checks were performed to remove corrupted samples, incomplete entries, and extremely small or noisy segmentations, using a largest-connected-component filter to maintain anatomical consistency.

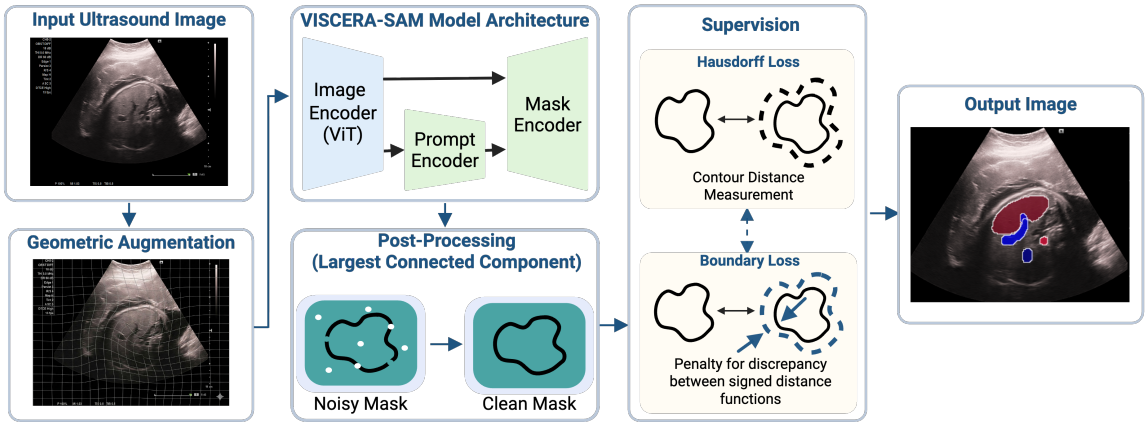


Figure 1: Overview of the VISCERA-SAM framework: Enhanced with boundary-aware optimization, Geometry-preserving augmentation, and LCC post-processing.

2.3. Model Architecture

UltraSAM (Meyer et al., 2025) is an ultrasound-specialized adaptation of the Segment Anything Model (SAM). To ensure robust generalization across sonographic textures, UltraSAM is pre-trained on the US-43d dataset, a large-scale collection of 282,321 image–mask pairs aggregated from 43 open-access datasets and covering 58 anatomical structures. This diversity helps the encoder learn domain-invariant ultrasound features prior to task-specific fine-tuning.

Given an input ultrasound image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$ and a set of interactive prompts P, the model predicts a set of segmentation masks $\hat{M}$ for anatomical instances of interest:

$$\hat{M} = f_\theta(I, P), \tag{1}$$

where f_θ denotes the UltraSAM network parameterized by θ. Throughout, we use $H_0 = W_0 = 1024$.

UltraSAM follows the SAM-style architecture and consists of four components: (i) an image encoder based on a ViT-B backbone (Dosovitskiy, 2020), (ii) a prompt encoder that maps point and box prompts into sparse and dense token embeddings, (iii) a lightweight two-way transformer decoder, and (iv) a mask prediction head with cascaded refinement.

2.3.1. Overall Framework

Let I denote the preprocessed ultrasound image and P the set of prompts (Sec. 2.3.3). The image encoder E_{img} produces a dense feature map, and the prompt encoder E_{prompt} produces sparse and dense prompt embeddings:

$$F_{\text{img}} = E_{\text{img}}(I) \in \mathbb{R}^{B \times C \times H \times W}, \tag{2}$$

$$F_{\text{sparse}}, F_{\text{dense}}, T = E_{\text{prompt}}(P), \tag{3}$$

where B is the batch size, $C = 256$ the encoder channel dimension, $(H, W) = (64, 64)$ the feature resolution, F_{sparse} sparse prompt embeddings, F_{dense} dense mask-related embeddings, and $T \in \mathbb{R}^{N_{\text{tok}} \times C}$ a set of learnable output tokens ($N_{\text{tok}} = 5$; three multi-mask tokens, one single-mask token, and one IoU token).

The transformer decoder D_{trans} performs two-way attention between image and prompt features and produces updated tokens and image features:

$$\tilde{T}, \tilde{F}_{\text{img}} = D_{\text{trans}}(F_{\text{img}}, F_{\text{sparse}}, F_{\text{dense}}, T). \tag{4}$$

Finally, the mask prediction head H_{mask} generates mask logits and IoU scores:

$$\hat{M}, \hat{\mathbf{s}} = H_{\text{mask}}(\tilde{T}, \tilde{F}_{\text{img}}), \tag{5}$$

$$\hat{M} = \{\hat{M}_j\}_{j=1}^{N_{\text{mask}}}, \quad N_{\text{mask}} = 4, \tag{6}$$

where $\hat{M}_j$ denotes the j-th predicted mask logit map and $\hat{\mathbf{s}} \in \mathbb{R}^{N_{\text{mask}}}$ the corresponding IoU scores.

The cascaded refinement module re-injects predicted masks as additional prompts (Sec. 2.3.6) to iteratively improve segmentation quality.

Training Strategy. We adopt **full fine-tuning** of all UltraSAM parameters rather than using adapter-based approaches (e.g., SAM-Med), due to the significant domain gap between natural images and fetal ultrasound (speckle noise, low contrast, and distinct texture statistics). We optimize a combination of region-based and boundary-aware losses (Sec. 2.4), including Dice and BCE for segmentation and a regression loss for IoU scores.

2.3.2. Image Encoder

Preprocessing. Ultrasound images are normalized following the SAM convention. Let $I_{\text{raw}} \in \mathbb{R}^{H_0 \times W_0 \times 3}$ be the original BGR image. We first convert to RGB and apply channel-wise affine normalization:

$$I_{\text{rgb}}(x, y, c) = \text{BGR2RGB}(I_{\text{raw}}(x, y, c)), \tag{7}$$

$$I(x, y, c) = \frac{I_{\text{rgb}}(x, y, c) - \mu_c}{\sigma_c}, \tag{8}$$

where $(\mu_c) = [123.675, 116.28, 103.53]$ and $(\sigma_c) = [58.395, 57.12, 57.375]$. Images are then padded so that $H_0 = W_0 = 1024$.

Patch embedding and transformer encoding. We use a ViT-B backbone with patch size $p = 16$. The input is partitioned into $N = (H_0/p)(W_0/p) = 4096$ non-overlapping patches:

$$\{P_i\}_{i=1}^{N} = \text{Patchify}(I), \qquad P_i \in \mathbb{R}^{p^2 \cdot 3}, \tag{9}$$

each linearly projected into a $d_{\text{vit}} = 768$-dimensional embedding. After adding 2D positional embeddings, the sequence is processed by $L = 12$ window-based transformer blocks with multi-head self-attention and FFNs, following SAM's encoder design.

The final token sequence is reshaped into a 2D feature map and projected to $C = 256$ channels:

$$F_{\text{img}} = \text{Reshape}(\mathbf{Z}_L)\mathbf{W}_p \in \mathbb{R}^{B \times 256 \times 64 \times 64}, \tag{10}$$

where $\mathbf{Z}_L \in \mathbb{R}^{N \times d_{\text{vit}}}$ is the final transformer output and $\mathbf{W}_p \in \mathbb{R}^{d_{\text{vit}} \times 256}$.

2.3.3. Prompt Encoder

The prompt encoder E_{prompt} maps interactive point and box prompts into sparse embeddings, dense mask embeddings, and learnable output tokens.

Point and box prompts. A point prompt is defined by its spatial coordinate (x, y) and a label $t \in \{\text{pos}, \text{neg}, \text{box}, \text{other}\}$. We encode (x, y) with a learned positional embedding $\mathbf{e}_{\text{pos}}(x, y)$ and the label with $\mathbf{e}_{\text{label}}(t)$, then project to the SAM embedding dimension $C = 256$:

$$\mathbf{e}_p = \mathbf{W}_p\big[\mathbf{e}_{\text{pos}}(x, y) \| \mathbf{e}_{\text{label}}(t)\big] \in \mathbb{R}^C. \tag{11}$$

Box prompts are represented by two corner points encoded in the same manner. During training, point and box prompts are sampled with equal probability to encourage robustness.

Per image, the sparse prompt embeddings F_{sparse} and dense prompt-aware embeddings F_{dense} are constructed at the feature resolution $(H, W) = (64, 64)$, together with $N_{\text{tok}} = 5$ learnable output tokens T.

2.3.4. Transformer Decoder

The transformer decoder D_{trans} is a lightweight two-layer stack of two-way attention blocks operating in the SAM feature space ($C = 256$). We flatten F_{img} into a sequence $\mathbf{X}^{\text{img}} \in \mathbb{R}^{B \times HW \times C}$ ($HW = 64 \times 64$) and broadcast the sparse prompt tokens and T to form $\mathbf{X}^{\text{sparse}}$.

Each two-way block alternates: (i) attention from sparse tokens to image tokens, and (ii) attention from image tokens back to sparse tokens, each followed by FFNs and residual connections. After two blocks, we obtain updated output tokens and image features, denoted $\tilde{T}$ and $\tilde{F}_{\text{img}}$.

2.3.5. Mask Prediction Head

The mask prediction head (`SAMHead`) converts decoder outputs into high-resolution masks using an upsampling path and hypernetwork-based dynamic convolution.

Upsampling path. The updated image features $\tilde{F}_{\text{img}} \in \mathbb{R}^{BN_{\text{inst}} \times C \times 64 \times 64}$ are upsampled by two transposed convolutions to obtain $U_2 \in \mathbb{R}^{BN_{\text{inst}} \times 32 \times 256 \times 256}$, which serves as the mask feature map.

Dynamic filters and IoU prediction. For each output token $\tilde{t}_j \in \mathbb{R}^C$ ($j = 1, \ldots, N_{\text{mask}}$), a small MLP produces a 32-dimensional vector $\mathbf{w}_j$ that acts as a dynamic 1×1 convolution on U_2:

$$\hat{M}_j = \sum_{c=1}^{32} w_{j,c} \cdot U_2^{(c)}, \qquad \hat{M}_j \in \mathbb{R}^{256 \times 256}. \tag{12}$$

A separate MLP predicts an IoU score $\hat{s}_j$ for each mask. During training, IoU predictions are supervised with mean squared error between $\hat{s}_j$ and the true IoU with the ground-truth mask M_{gt}.

Segmentation loss. The final selected mask $\hat{M}$ is supervised using a combination of binary cross-entropy (BCE) and soft Dice loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\Omega|}\sum_{x\in\Omega}\Big[M_{\text{gt}}(x)\log\sigma(\hat{M}(x)) + (1-M_{\text{gt}}(x))\log(1-\sigma(\hat{M}(x)))\Big], \tag{13}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{x\in\Omega}\sigma(\hat{M}(x))M_{\text{gt}}(x)+\epsilon}{\sum_{x\in\Omega}\sigma(\hat{M}(x))+\sum_{x\in\Omega}M_{\text{gt}}(x)+\epsilon}, \tag{14}$$

where σ is the sigmoid function, Ω the pixel set, and ϵ a small constant. The overall segmentation loss is

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}}, \tag{15}$$

and the total loss combines segmentation and IoU regression:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}}. \tag{16}$$

2.3.6. Cascaded Mask Refinement

UltraSAM implements a single-step cascaded refinement mechanism that iteratively updates prompts using the previously predicted mask.

Given mask logits $\hat{M}^{(t)}$ at iteration t, we obtain a soft mask

$$\tilde{M}^{(t)} = \sigma(\hat{M}^{(t)}), \tag{17}$$

encode it with a small CNN into a dense mask feature map $F^{(t)}_{\text{mask}} \in \mathbb{R}^{256\times64\times64}$ aligned with F_{img}, and derive a bounding box

$$b^{(t)} = \text{BBox}(\tilde{M}^{(t)}), \tag{18}$$

which replaces the original point prompts for that instance. At refinement iteration $t+1$, the model is run again with updated prompts and mask features:

$$\hat{M}^{(t+1)}, \hat{\mathbf{s}}^{(t+1)} = f_\theta(I, P^{(t+1)}, F^{(t)}_{\text{mask}}), \tag{19}$$

and we use $T = 1$ refinement iteration during both training and inference:

$$\hat{M}^{\text{final}} = \hat{M}^{(T)}. \tag{20}$$

2.4. Enhancements via Post-Processing and Boundary-Aware Loss Optimization

To improve the structural consistency and robustness of UltraSAM predictions on ultrasound data, we introduce two complementary modifications: (i) a post-processing step based on the *Largest Connected Component* (LCC) criterion, and (ii) an enhanced loss function that integrates boundary-sensitive objectives, including Boundary Loss and a Hausdorff Distance-based loss. These modifications target typical ultrasound failure modes such as speckle-induced artifacts, fragmented boundaries, and outlier regions that disproportionately affect distance-based metrics (Karimi and Salcudean, 2019).

2.4.1. Largest Connected Component (LCC) Post-Processing

Ultrasound predictions often contain small isolated artifacts caused by speckle noise or local activation spikes (Wang, 2020; Noble and Boukerroui, 2006). These artifacts may not significantly affect region-based metrics but can produce large deviations in contour-based metrics such as the Hausdorff distance (Huttenlocher et al., 2002). To suppress these outlier regions, we apply an LCC filter to each predicted mask.

Let $\mathcal{C}(M)$ denote the set of connected components of a predicted mask M:

$$\mathcal{C}(M) = \{C_1, C_2, \ldots, C_K\}.$$

We retain only the dominant component:

$$M_{\text{LCC}} = \arg \max_{C_i \in \mathcal{C}(M)} |C_i|.$$

This simple post-processing step removes distant noisy fragments while preserving the primary anatomical structure and empirically reduces Hausdorff error by eliminating isolated false positives that dominate worst-case boundary deviations.

2.4.2. Boundary-Aware Loss Optimization

The baseline training regime uses region-based loss terms such as Dice Loss and Binary Cross-Entropy (BCE). While sufficient for coarse mask accuracy, these losses provide limited supervision for boundary alignment, which is a critical factor in ultrasound images where object contours are often faint, irregular, or partially missing.

To address this limitation, we incorporate two additional loss terms: the Boundary Loss and a differentiable approximation of the Hausdorff Distance Loss.

Boundary Loss. Boundary Loss aligns the predicted and ground-truth level sets by penalizing discrepancies between their signed distance functions (SDFs). Let ϕ_{gt} and ϕ_{pred} denote the SDFs of the ground-truth and predicted masks, respectively. The loss is computed as:

$$\mathcal{L}_{\text{boundary}} = \int_{\Omega} |\phi_{\text{gt}}(x) - \phi_{\text{pred}}(x)| \, dx.$$

This formulation directly supervises contour placement, providing stronger gradients along the object's perimeter than region-based losses.

Hausdorff Distance Loss. The Hausdorff distance quantifies the maximum boundary deviation:

$$d_H(\Gamma_{\text{pred}}, \Gamma_{\text{gt}}) = \max \left\{ \sup_{x \in \Gamma_{\text{pred}}} \inf_{y \in \Gamma_{\text{gt}}} \|x - y\|, \sup_{y \in \Gamma_{\text{gt}}} \inf_{x \in \Gamma_{\text{pred}}} \|x - y\| \right\}.$$

We adopt a differentiable surrogate inspired by the generalized mean:

$$\mathcal{L}_H = \left(\frac{1}{|\Gamma_{\text{pred}}|} \sum_{x \in \Gamma_{\text{pred}}} d(x, \Gamma_{\text{gt}})^p \right)^{1/p} + \left(\frac{1}{|\Gamma_{\text{gt}}|} \sum_{y \in \Gamma_{\text{gt}}} d(y, \Gamma_{\text{pred}})^p \right)^{1/p}, \qquad p = 2.$$

Table 1: Quantitative comparison before and after applying the proposed enhancements.

	(a) Baseline UltraSAM					(b) Enhanced UltraSAM				
Organ	Bbox	SegmAP	Dice	mIoU	HD	Bbox	SegmAP	Dice	mIoU	HD
Artery	0.614	0.500	0.8614	0.7586	6.21	0.663	0.554	0.8766	0.7821	5.48
Liver	0.891	0.729	0.9214	0.8555	21.41	0.908	0.839	0.9492	0.9039	12.92
Stomach	0.811	0.645	0.9002	0.8230	9.98	0.798	0.691	0.9175	0.8515	7.47
Vein	0.794	0.484	0.8535	0.7483	10.82	0.792	0.594	0.8876	0.8001	7.90
Mean	**0.778**	**0.589**	**0.8841**	**0.7963**	**12.11**	**0.790**	**0.670**	**0.9100**	**0.8300**	**8.44**

Final Loss. The full training objective is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{BL}}\mathcal{L}_{\text{boundary}} + \lambda_H\mathcal{L}_H.$$

This hybrid loss encourages: (i) region-level overlap, (ii) accurate contour geometry, and (iii) robustness to extreme outlier deviations.

2.4.3. Geometric Data Augmentation

Fetal veins, especially the intrahepatic umbilical vein, show substantial shape variability due to fetal positioning and compression, which standard affine transforms cannot fully model. To improve robustness to such non-rigid deformations, we incorporate targeted geometric augmentations, including elastic deformation to simulate tissue compression and grid distortion to introduce twisting or stretching effects. These operations encourage the model to handle irregular venous boundaries and prevent overfitting to the idealized, circular vessel shapes seen in canonical cross-sections.

3. Quantitative Results

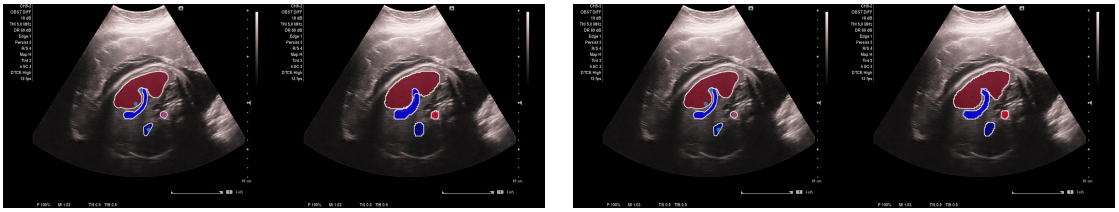
Tables 1 report performance before and after applying the proposed enhancements. The updated model demonstrates consistent improvements across all organs, particularly in SegmAP, Dice, and Hausdorff metrics.

The enhanced configuration reduces the average Hausdorff error by 30–40%, improves SegmAP by 13.7%, and increases Dice by 2.5-3% on average, demonstrating the effectiveness of boundary-aware optimization and post-processing refinements.

4. Discussion

4.1. Why Ultrasound Benefits From Boundary-Aware Supervision

Ultrasound segmentation is uniquely challenged by speckle noise, acoustic shadowing, low contrast, and partially missing boundaries (Shan et al., 2012; Noble and Boukerroui, 2006; Wang, 2020). Unlike CT or MRI, where edges are well defined, ultrasound often provides weak gradients for region-based objectives such as Dice or BCE (Taha and Hanbury, 2015; Wang et al., 2018). Boundary-aware losses directly supervise contour geometry (Kervadec et al., 2019), leading to sharper delineation and more stable optimization. Complementarily, Largest Connected Component (LCC) filtering removes isolated false positives that



(*a*) Baseline model prediction. (*b*) Improved model prediction.

Figure 2: Qualitative analysis of baseline UltraSAM (a) and our VISCERA-SAM (b).

disproportionately inflate distance-based metrics like Hausdorff. Together, these components yield smoother, more anatomically coherent predictions that better support clinical interpretation.

4.2. Qualitative and Clinical Relevance

As illustrated in Fig. 2, VISCERA-SAM produces more continuous vessel walls and cleaner liver and stomach boundaries, especially in regions affected by attenuation or shadowing. These improvements are clinically meaningful: accurate depiction of venous morphology supports assessment of vascular development, while stable liver and stomach contours contribute to more reproducible abdominal circumference and volumetric measurements. Because small contour errors can propagate into significant deviations in biometric indices, the enhanced boundary fidelity of VISCERA-SAM strengthens its suitability for routine obstetric imaging.

5. Conclusion

We introduced **VISCERA-SAM**, a boundary-aware adaptation of UltraSAM tailored for the geometric challenges of fetal abdominal ultrasound. Although foundation models offer strong generalization, their performance in sonography remains limited by speckle noise, weak anatomical boundaries, and catastrophic forgetting (Chen et al., 2020; Shan et al., 2012). VISCERA-SAM addresses these issues through transfer learning combined with Boundary Loss, Hausdorff Loss, and Largest Connected Component refinement.

Experiments show that VISCERA-SAM markedly improves segmentation quality, achieving a mean Dice of 0.910 and reducing mean Hausdorff distance by nearly 30% (from 12.11 to 8.44). These gains in boundary fidelity are essential for accurate fetal abdominal circumference measurement, directly supporting screening for growth restriction and macrosomia. By producing anatomically coherent segmentations of the aorta, umbilical vein, stomach, and liver, VISCERA-SAM advances the reliability of automated prenatal biometry.

Data Availability

The fetal abdominal segmentation dataset used in this study is publicly available on Mendeley Data under a CC BY 4.0 license: https://doi.org/10.17632/4gcpm9dsc3.1. No additional restrictions apply. All experiments were performed exclusively on this dataset.

Code Availability

The code used for model training and evaluation will be released upon publication at: https://github.com/john-minhle/VISCERA-SAM

Funding

This research received no external funding. Alternatively, add grant information here if applicable.

Conflict of Interest

The authors declare that they have no conflicts of interest related to this work.

References

Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, and Wen Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4):83–93, 2020.

C Da et al. Fetal abdominal structures segmentation dataset using ultrasonic images. *Mendeley Data*, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Frank P Hadlock, RB Harrist, RS Sharman, RL Deter, and SK Park. Estimation of fetal weight with the use of head, body, and femur measurements: a prospective study. *American Journal of Obstetrics and Gynecology*, 151(3):333–337, 1985.

Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 2002.

Davood Karimi and Septimiu E Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2):499–513, 2019.

Hoel Kervadec, Jihed Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 285–296. PMLR, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Adrien Meyer, Aditya Murali, Farahdiba Zarin, Didier Mutter, and Nicolas Padoy. Ultrasam: a foundation model for ultrasound using large open-access segmentation datasets. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–10, 2025.

J Alison Noble and Djamal Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transactions on Medical Imaging*, 25(8):987–1010, 2006.

LJ Salomon, Z Alfirevic, F Da Silva Costa, RL Deter, F Figueras, T Ghi, P Glanc, A Khalil, W Lee, R Napolitano, et al. Isuog practice guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound in Obstetrics & Gynecology*, 53(6):715–723, 2019.

Juan Shan, HD Cheng, and Yuxuan Wang. Completely automated segmentation approach for breast ultrasound images using multiple-domain features. *Ultrasound in medicine & biology*, 38(2):262–275, 2012.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):29, 2015.

Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018.

Ziyang Wang. Deep learning in medical ultrasound image segmentation: A review. *arXiv preprint arXiv:2002.07703*, 2020.

Appendix A. Appendix: Expanded Discussion

A.1. Organ-Specific Effects of Boundary-Aware Supervision

Artery. Arterial walls are thin and low-contrast, causing the baseline to produce fragmented false positives that inflate Hausdorff distance. Boundary and Hausdorff losses enforced smoother vessel walls, while LCC filtering removed isolated noise, improving Dice and reducing Hausdorff error by 12%.

Liver. Dice increased from 0.9214 to 0.9492 and mIoU from 0.8555 to 0.9039. Boundary-aware supervision corrected posterior contour irregularities caused by attenuation, producing smoother liver capsules relevant for volume estimation and lesion monitoring.

Stomach. The baseline often misclassified gas or fluid as boundaries. Hausdorff Loss reduced extreme contour deviations, improving Dice from 0.9002 to 0.9175 and lowering Hausdorff distance by over 25%.

Vein. Dice improved from 0.8535 to 0.8876 and Hausdorff from 10.82 to 7.90. The enhanced model better captured venous morphology and continuity while suppressing speckle-induced false positives.

Overall Impact. Boundary-aware training consistently improved SegmAP and reduced mean Hausdorff distance from 12.11 to 8.44, yielding more anatomically faithful boundaries that benefit downstream measurement and surgical planning.

A.2. Additional Qualitative Analysis

Figures 2(a) and 2(b) highlight the gap between UltraSAM and VISCERA-SAM. The baseline exhibits fragmented vessel predictions, collapsed venous structures, and unstable liver and stomach boundaries. VISCERA-SAM produces smoother, more coherent contours, maintaining venous caliber and producing stable liver and stomach outlines even under shadowing. These refinements improve robustness in low-contrast, speckled regions and support more reliable biometric estimation.

A.3. Limitations and Future Work

This study is restricted to a single-center dataset of term pregnancies, which may limit generalizability to earlier gestational ages or broader populations. Although images originated from several ultrasound machines, the dataset does not capture the full variability of global imaging equipment or operator technique. Additionally, our evaluation is limited to 2D standard AC-plane images; extending the framework to 3D ultrasound or real-time cine imaging remains a promising direction. Multi-center prospective validation is needed to establish clinical robustness.