Computer Methods and Programs in Biomedicine 275 (2026) 109178

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Adapt or specialize? A comprehensive evaluation of adapted SAM versus task-specific CNNs for fetal abdominal segmentation

Maria Chiara Fiorentino [a], Lorenzo Federici [a,*], Alessandro Pietro La Camera [b], Enrico Gianluca Caiani [b]

[a] Department of Information Engineering, Marche Polytechnic University, Ancona, Italy
[b] Department of Electronics, Information and Bioengineering Politecnico di Milano, Milano, Italy

ARTICLE INFO

Keywords:
Fetal ultrasound
Deep learning
Abdomen
Segmentation

ABSTRACT

Background: The fetal abdomen is crucial in prenatal screening, offering key insights into fetal growth and congenital anomalies. However, segmenting internal abdominal structures in ultrasound (US) remains challenging due to anatomical variability, overlapping organs, and low contrast. While CNN-based models have shown strong performance in fetal head analysis, most existing methods focus on biometric measurements (e.g., head or abdominal circumference), leaving internal abdominal organ segmentation largely underexplored. Recently, foundation models like the Segment Anything Model (SAM) have emerged as flexible alternatives, enabling zero- or few-shot segmentation. Yet, their performance on fetal US remains poorly understood, and the need for adaptation is still an open question.

Methods: We compare two segmentation strategies: (1) task-specific CNNs, including UNet, Attention UNet, nnUNet, DeepLabv3+, and their focal-loss variants; and (2) adapted SAM-based models. The latter includes zero-shot variants (SAMPoint, SAMBBox), pre-trained models (SAM Med2DBBox, MedSAMBBox), and adapted configurations, including lightweight fine-tuned models (SAM-LoRA, MedSAM-FrozenEncoder) and SAM Med2D variants with adapted layers. Experiments are conducted on a curated dataset of fetal abdominal US images with manual segmentations of the liver, stomach, artery, and umbilical vein. Performance is evaluated using Dice Similarity Coefficient, Intersection over Union, and precision. Statistical significance is assessed via pairwise Friedman chi-square tests.

Results & Conclusions: Zero-shot SAM variants performed poorly, particularly on small or low-contrast structures. In contrast, adapted SAM models consistently outperformed CNNs, reaching DSC scores up to 0.90 (liver) and 0.80 (artery). Prompt-based interaction enables semi-automated, human-in-the-loop workflows, supporting clinical applicability.

1. Introduction

Fetal ultrasound (US) is the primary imaging modality for prenatal screening, offering a non-invasive and real-time assessment of fetal development [1]. Among the various anatomical regions analyzed during routine examinations, the fetal abdomen provides critical information for evaluating growth patterns, detecting congenital anomalies, and assessing organ function [2]. Key structures such as the stomach, liver, kidneys, and bowel are routinely examined to monitor fetal well-being and detect potential pathological conditions, including gastrointestinal atresia, renal agenesis, and hepatomegaly [3]. This makes the accurate identification of organ shapes essential for the proper assessment of fetal development [4]. However, the manual segmentation of fetal abdomen structures is time-consuming and operator-dependent, making it challenging to ensure consistency and integrate into high-throughput clinical workflows. These limitations have driven growing interest in automated segmentation methods that can enhance the accuracy, efficiency, and consistency of fetal US analysis [5,6].

In recent years, deep learning (DL) has revolutionized medical image analysis, offering state-of-the-art performance in tasks such as classification, detection, and segmentation. Convolutional neural networks (CNNs), in particular, have demonstrated exceptional accuracy in delineating anatomical structures across various imaging modalities, outperforming traditional image processing techniques [7,8]. These models learn hierarchical features directly from the data, enabling precise and robust segmentation even in challenging imaging scenarios with limited data, such as US [9]. Most research efforts in fetal US

* Corresponding author.
E-mail address: s1125147@pm.univpm.it (L. Federici).

https://doi.org/10.1016/j.cmpb.2025.109178
Received 1 July 2025; Received in revised form 9 November 2025; Accepted 20 November 2025
Available online 26 November 2025

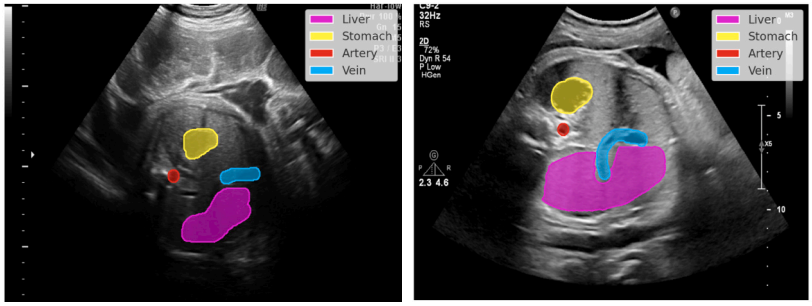


Fig. 1. Example of fetal abdominal ultrasound images with segmentations of key anatomical structures. The highlighted regions correspond to the arteries, liver, stomach, and veins. These structures are critical for assessing fetal development and detecting potential abnormalities.

have focused on the segmentation of the fetal head, where CNN-based models have demonstrated high accuracy and robustness [1]. In fetal head segmentation, several studies have leveraged UNet to automatically delineate cranial structures for biometric measurements and anomaly detection. U-Net, with its symmetric encoder–decoder architecture, has been widely adopted due to its ability to learn hierarchical features at multiple levels [10], especially when training data is limited [11]. The encoder captures low-level features, rich in spatial details, while the decoder extracts high-level semantic information, essential for accurate segmentation. However, a direct concatenation of these features does not always yield optimal performance, as low- and high-level information differ significantly in representation. To address this limitation, researchers have proposed numerous enhancements to the UNet framework, incorporating techniques such as attention mechanisms [2,12,13], residual connections [14], multi-scale feature fusion [15] and squeeze atrous pooling [11]. Regarding the fetal abdomen, UNet has also been employed for estimating the abdominal circumference [16]. However, no studies have yet explored the segmentation of internal fetal structures, which presents a significantly greater challenge compared to fetal head or abdominal circumference segmentation. This is due to anatomical variability, differences in tissue echogenicity, frequent overlaps between adjacent organs, and the fact that even within the same axial plane, the spatial configuration of organs may vary considerably due to fetal movements and subtle changes in probe orientation (see Fig. 1).

For such complex tasks, recent research has shown a growing interest in more flexible and generalizable approaches, such as foundation models, which have demonstrated strong cross-domain transfer capabilities and the potential to operate effectively with limited annotated data [17]. This trend has recently extended to the fetal US domain, as demonstrated by [18], where foundation models have been adapted to fetal imaging to enhance generalizability across anatomical districts. Building upon this momentum, new architectures specifically designed for segmentation have also emerged, such as the Segment Anything Model (SAM) developed by Meta AI [19]. Unlike conventional DL models, SAM leverages a prompt-based segmentation strategy, allowing users to interactively refine segmentations through simple inputs—such as points, bounding boxes, or masks. This flexibility makes it particularly suitable for medical imaging applications where labeled data are scarce and where intra-class variability is significant. Applying SAM to fetal abdomen segmentation could provide a zero-shot or few-shot segmentation capability, enabling the identification of complex anatomical structures without the need for extensive retraining on domain-specific datasets. However, despite its potential, SAM's performance in the fetal US remains largely unexplored, as is the extent to which adaptations may be required for effective application. Moreover, its effectiveness compared to specialized CNN architectures, which have been meticulously optimized for fetal US analysis, has yet to be thoroughly investigated.

Inspired by these considerations, our aim is to address two critical questions:

- Is SAM truly superior to specialized CNNs for fetal abdomen segmentation, especially in the context of limited data, as is common in this field? Does it require adaptation?
- Which approach offers the best trade-off between accuracy and clinical applicability?

To address these questions, this study will use a fetal abdominal US image dataset with segmentations of key anatomical structures, including the arteries, liver, stomach, and veins. Our main contribution can be summarized as follows:

1. We introduce the first dedicated framework for segmenting internal fetal abdominal structures, extending beyond the conventional focus on abdominal circumference estimation. This framework establishes a benchmark for the automated segmentation of key organs and vascular structures in fetal US, addressing the unique challenges posed by anatomical variability and imaging artifacts.
2. We conduct a rigorous evaluation to determine whether SAM can outperform a state-of-the-art specialized CNN in fetal abdominal segmentation. To ensure a fair and comprehensive comparison, we leverage CNN architectures by integrating the latest advancements in DL, such as attention mechanisms, multi-scale feature fusion, and residual connections, while simultaneously identifying the optimal adaptation strategy for SAM to enhance its performance in fetal US segmentation.

2. Methods

This study aims to provide valuable insights into the comparative efficacy of CNNs and SAM for fetal abdomen structure segmentation (see Fig. 2). In Section 2.1, we introduce the network architectures and explain the rationale behind their selection. Section 2.2 outlines the data used for evaluation. Finally, the training procedure and performance evaluation are presented in Sections 2.3 and 2.4, respectively.

2.1. Network architectures

2.1.1. Adapted SAM

SAM [19] is the first promptable foundation model for image segmentation. It is trained on the large-scale SA-1B dataset, which contains an unprecedented number of images and annotations, enabling remarkable zero-shot generalization capabilities. SAM employs a transformer-based architecture [20], a paradigm that has demonstrated exceptional performance in natural language processing [21] and image recognition tasks [22]. Specifically, SAM consists of three main components:

1. Image Encoder: A Vision Transformer (ViT)-based encoder [22], pre-trained using the Masked Autoencoder (MAE) framework [23], designed to efficiently process high-resolution images. It takes a 1024×1024 image as input and outputs a 64×64 downscaled feature map.

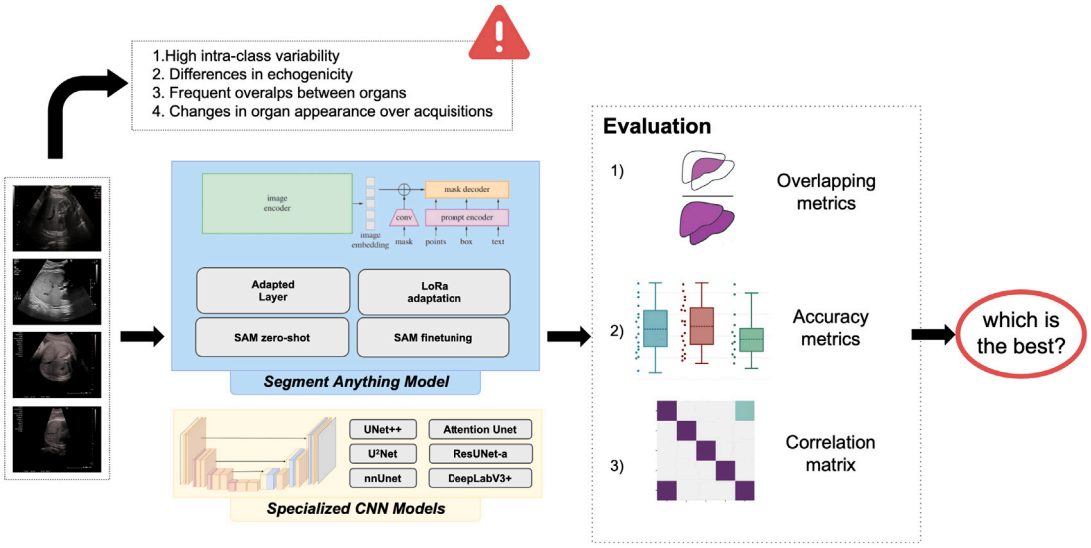


Fig. 2. An overview of the proposed pipeline for fetal abdominal structure segmentation in ultrasound (US) images. The input consists of 2D US scans of the fetal abdomen. We evaluate two segmentation strategies: (i) the Segment Anything Model (SAM), enhanced with domain-specific adaptation via LoRA and adapted layers; and (ii) specialized convolutional neural networks (CNNs), including UNet++, Attention UNet, nnUNet, and others. Performance is compared across models using DICE and IoU metrics along with Precision and correlation matrix.

2. Prompt Encoder: SAM supports both sparse (e.g., points, boxes) and dense (e.g., masks) prompts. Sparse prompts are encoded using positional encoding [24] combined with learnable embeddings: foreground and background points are represented by two learnable tokens, while bounding boxes are encoded based on the coordinates of their top-left and bottom-right corners. Dense mask prompts, which match the spatial resolution of the input image, are embedded using convolutional (conv) layers and summed element-wise with the image embedding.
3. Mask Decoder: A lightweight module composed of two transformer layers, a dynamic mask prediction head, and an Intersection-over-Union (IoU) score regression head. The mask prediction head generates three masks corresponding to the whole object, a part of the object, and a subpart.

This architecture was adapted in SAM-Med by integrating an Encoder Adapter, a lightweight module within the image encoder, designed to enhance the model's ability to capture medical image-specific features [25]. The Encoder Adapter operates by refining feature representations at both the channel and spatial levels. For the channel dimension, the adapter first applies global average pooling to compress the resolution of the input feature map to a compact representation of size $C \times 1 \times 1$. This compressed representation is then passed through a linear layer that reduces the number of channels, followed by another linear layer that restores them, using a compression ratio of 0.25 to efficiently retain relevant information. The output is then passed through a sigmoid activation function, producing a set of weights that are multiplied element-wise with the original feature map. This operation enhances the most relevant features while suppressing less informative ones, refining the model's ability to focus on important structures. For the spatial dimension, the adapter downscales the spatial resolution of the feature map using a convolutional layer with a stride of two. The reduced feature map is then upsampled using a transposed convolution, ensuring that the final output maintains the same spatial resolution and number of channels as the original input. A skip connection is added after each adapter layer.

Another architectural modification was introduced in [26], where the authors leveraged *LoRA (Low-Rank Adaptation)* to efficiently fine-tune the image encoder [27]. Within the image encoder, LoRA is applied to the self-attention layers that compute the *query (Q)*, *key (K)*, and *value (V)* representations. Typically, these are computed using a single linear transformation:

$$\mathrm{QKV}(X) = XW_{\mathrm{qkv}},$$

where $X \in \mathbb{R}^{B\times N\times d}$ is the input sequence, and $W_{\mathrm{qkv}} \in \mathbb{R}^{d\times 3d}$ is the weight matrix projecting the input to Q, K, and V representations. In the LoRA framework, the original projection is preserved and a low-rank residual is added to the query and value paths:

$$\mathrm{Q}_{\mathrm{adapted}} = XW_q + XAB, \quad \mathrm{V}_{\mathrm{adapted}} = XW_v + XA'B',$$

where $A, A' \in \mathbb{R}^{d\times r}$, $B, B' \in \mathbb{R}^{r\times d}$, and $r \ll d$ is the rank of the decomposition. The key projection W_k remains unchanged to reduce overhead. During training, only the parameters A, B, A', B' are updated, while the original weights remain frozen.

To determine the optimal configuration, we conducted a series of experiments with different adaptation strategies:

1. Zero-shot evaluation: First, we evaluated SAM's zero-shot segmentation capability by applying it directly to our dataset without fine-tuning. We conducted this assessment using both the original weights trained on natural images, SAM-Med 2D [25] and MedSam [28] weights.
2. SAM-Med-inspired Adaptation: Motivated by [25], we adopted a hybrid fine-tuning approach. We froze the image encoder to retain the general visual representation learned from large-scale natural images while training the learnable adapter layers in each Transformer block to capture domain-specific features. The prompt encoder was fine-tuned using both point prompts (*Adapted SAM Med2D*$_{\mathrm{Points}}$) and bounding box prompts (*Adapted SAM Med2D*$_{\mathrm{BBox}}$). Additionally, we updated the parameters of the mask decoder through interactive training, starting from SAM-Med's weights to leverage prior knowledge acquired from the largest medical image dataset.
3. SAM Lora-Adaptation: we employed LoRA [29] specifically on the image encoder. This choice allowed us to efficiently fine-tune the encoder by adapting its representations to the medical imaging domain without significantly increasing computational requirements (*SamLora*). At the same time, both the prompt encoder — which exploited bounding box annotations as a strong prior to facilitate convergence — and the mask decoder were fully trainable, enabling a comprehensive adaptation of SAM to the specific task of fetal abdomen segmentation [26].

4. SAM Finetuning: we performed the fine-tuning of the SAM model starting from the pre-trained MedSAM weights [28], in order to facilitate faster and more stable convergence during training. The image encoder was kept frozen to preserve the general visual representations learned during large-scale pretraining. This choice was motivated by the aim to prevent overfitting — particularly in the presence of our limited data — and to concentrate the optimization on the more task-specific components (i.e., prompt encoder and mask decoder) responsible for segmentation ($MedSamFrozeEncoder$).

2.1.2. CNNs

A comparative analysis of CNN architectures is conducted to evaluate their effectiveness in fetal abdomen segmentation. The selected architectures — U-Net [30], U-Net++ [31], Attention U-Net [32], U^2-Net [33], ResUnet-a [34],nnUnet [35] DeepLabV3+ [36] and EMCAD [37] — were chosen, as mostly used in [38], due to their distinct advantages in medical image segmentation.

U-Net: A well-established architecture known for its symmetric encoder–decoder structure and skip connections, which facilitate precise localization by preserving spatial information. The skip connections help recover fine details lost during downsampling, making U-Net particularly effective for medical image segmentation tasks.

U-Net++: An extension of U-Net designed to improve segmentation accuracy through a more refined skip connection mechanism. By introducing nested and dense skip pathways, U-Net++ enhances feature fusion, allowing for more precise boundary delineation and increased robustness to variations in image quality.

Attention U-Net: It incorporates an attention mechanism to selectively highlight relevant regions of the input image while suppressing background noise. This enables the model to focus on structures of interest, improving segmentation performance, particularly in cases where the anatomical boundaries are ambiguous or occluded.

U^2-Net: A two-stage nested U-Net architecture that employs a deeply supervised encoder–decoder framework with recurrent residual U-blocks. U^2-Net captures fine-grained details while maintaining a broad contextual understanding. Its hierarchical design ensures robust feature representation at multiple scales, improving segmentation accuracy even in challenging scenarios with complex backgrounds.

ResUnet-a: A residual U-Net architecture that integrates atrous (dilated) convolutions and a multi-scale context module (PSPooling). ResUnet-a enhances feature propagation through residual connections while leveraging atrous convolutions to expand the receptive field without increasing the number of parameters. Additionally, the inclusion of multi-scale pooling enables the network to better capture context and structural variations, making it particularly effective in handling scale variations and fine structures common in fetal US segmentation.

DeepLabV3+: Leverages an advanced feature extraction approach through atrous spatial pyramid pooling, which captures multiscale contextual information. The inclusion of an improved decoder module enhances spatial resolution, making DeepLabV3+ particularly effective for segmenting complex anatomical structures with varying shapes and sizes.

EMCAD: An efficient segmentation architecture that introduces a Multi-scale Convolutional Attention Decoding mechanism to enhance feature representation during upsampling. By combining multi-scale convolutional blocks with attention-based refinement, Emcad improves the preservation of fine anatomical boundaries while maintaining computational efficiency.

To maintain consistency with the original implementations and given the architectural complexity of the models, we selected U^2-Net, ResUnet-a, and nnU-Net as default architectures and trained them from scratch. Each of the other models was implemented with a ResNet50 backbone, initialized with ImageNet weights to mitigate overfitting and leverage pre-trained features. The models are trained using a final sigmoid activation function, allowing for the segmentation of multiple overlapping anatomical structures, as illustrated in Fig. 1.

2.2. Dataset

We used a publicly available dataset of fetal abdomen US images,[1] which was collected between September 2021 and September 2023. This dataset comprises US images from 169 subjects, each contributing a variable number of fetal AC images, amounting to nearly 1500 images.

Eligible participants were term pregnant women aged 18 years or older, either in labor or scheduled for delivery at the Maternity of University Hospital Polydoro Ernani de São Thiago in Florianópolis, Santa Catarina, Brazil. This included patients undergoing labor induction, cesarean section, or experiencing pregnancy complications such as membrane rupture, gestational diabetes, pre-eclampsia, or intrauterine growth restriction. Non-eligible cases comprised preterm pregnancies, multiple pregnancies, and pregnancies with fetal structural or chromosomal anomalies. Gestational age for all pregnancies was determined based on the earliest ultrasound examination.

The study received ethical approval from the Ethics Committee in Human Research at the authors' institution, Universidade Federal de Santa Catarina (UFSC), under approval number 4.971.754. All eligible participants were informed about the study objectives and provided written consent.

US images were acquired following a standardized protocol by expert clinicians. A common axial section of the fetal abdomen was captured, focusing on the widest part spanning the liver and including anatomical landmarks such as the stomach, aorta, spine, and the intrahepatic portion of the umbilical vein. The imaging was performed using Siemens Acuson, Voluson 730 (GE Healthcare US), and Philips-EPIQ Elite (Philips Healthcare US) systems, equipped with 2–9 MHz curved linear transducers. Post-processing features such as image smoothing, zoom, calipers, pointers, and Doppler measurements were disabled to ensure consistency. Tissue harmonic imaging was allowed, while frequency, gain, and focus adjustments were at the discretion of the performing physician. Images were digitally stored in DICOM format for offline analysis. A rigorous quality control process, conducted by two team members, ensured that images containing calipers or significant acoustic shadows from bony structures were excluded. Only images meeting strict quality standards were selected for further analysis.

For segmentation and anatomical assessment, expert clinicians utilized *3D Slicer* (version 5.2.2), a specialized medical imaging software, to delineate key abdominal structures, particularly the abdominal aorta, intrahepatic umbilical vein, stomach, and liver, ensuring precise annotations.

[1] https://data.mendeley.com/datasets/4gcpm9dsc3/1

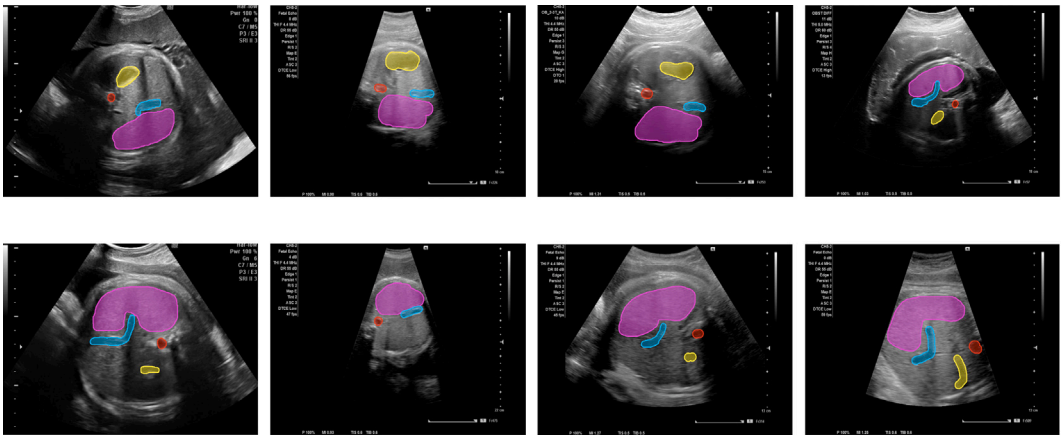


Fig. 3. Example images from the test set, showing the segmentation of key anatomical structures: artery, liver, stomach, and vein. These structures exhibit drastic shape variations across images, posing a challenge for the model's generalization. Additional challenges include differences in echogenicity, variations in size, and positional shifts of the anatomical structures.

To evaluate model performance, the dataset was split into training, validation, and test sets, ensuring that images from the same patient were not included in multiple subsets. Specifically, 60% of the subjects were allocated to the training set, 25% to the validation set, and 15% to the test set. This patient-level separation was crucial to prevent overoptimistic performance estimates, which could arise if images from the same patient were included across different subsets.

The test set was designed to include particularly challenging cases the model had to generalize to (see Fig. 3).

2.3. Training/inference settings

Before training, images are preprocessed through pixel intensity normalization. For CNN architectures, resizing is applied to 384×288 to ensure consistency across models. However, due to framework constraints, SAM requires resizing to 256×256. Additionally, on-the-fly data augmentation is applied to enhance dataset variability and improve generalization. This includes random rotations within $[-20°, +20°]$ ($p = 0.6$), horizontal flips ($p = 0.5$), and elastic deformations ($p = 0.4$), a strategy widely adopted in the literature. The objective of training is to minimize a loss function that quantifies the discrepancy between the predicted segmentation mask $\hat{Y}$ and the ground truth Y.

For CNN-based models, the primary loss function used is the Dice loss, which is well-suited for medical image segmentation due to its robustness in handling class imbalance:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum Y\hat{Y} + \epsilon}{\sum Y + \sum \hat{Y} + \epsilon} \tag{1}$$

where ϵ is a small constant for numerical stability.

Additionally, we employed the Focal Loss to further address class imbalance and improve segmentation performance in challenging cases. The Focal Loss is formulated as:

$$\mathcal{L}_{\text{Focal}} = -\alpha(1-\hat{Y})^{\gamma} Y \log(\hat{Y}) - (1-\alpha)\hat{Y}^{\gamma}(1-Y)\log(1-\hat{Y}) \tag{2}$$

where α is a weighting factor to balance class contributions, and γ is a focusing parameter that reduces the weight of well-classified examples to focus more on hard-to-classify instances.

Training is conducted using a batch size of 8, an Adam optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a Smoothed ReduceLROnPlateau learning rate scheduler (monitoring `val_loss`, factor=0.2, patience=3, min_lr=1×10^{-6}, window_size=4). The initial learning rate is set to 1×10^{-4}, and models are trained for 50 epochs. Additionally, Smoothed Early Stopping (monitoring `val_loss`, patience=8, window_size=4) is applied to prevent overfitting.

For the adapted SAM models, the same training settings are used, except for the loss function, which is set to the default: a combination of focal loss and Dice loss. In this study, point and bounding box prompts were automatically generated from the ground truth annotations to mimic the actions that a clinician would perform during the segmentation process.

2.4. Performance evaluation

To evaluate segmentation performance, we consider three key metrics: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and Precision. The DSC measures the overlap between the predicted segmentation mask $\hat{Y}$ and the ground truth mask Y, and is defined as:

$$DSC = \frac{2\sum Y\hat{Y}}{\sum Y + \sum \hat{Y}} \tag{3}$$

A higher DSC indicates better segmentation performance. The IoU quantifies the ratio between the intersection and the union of the predicted and ground truth masks:

$$IoU = \frac{\sum Y\hat{Y}}{\sum Y + \sum \hat{Y} - \sum Y\hat{Y}} \tag{4}$$

This metric provides a stricter evaluation than DSC, as it penalizes false positives and false negatives more strongly.

Precision assesses the proportion of correctly predicted positive pixels relative to all predicted positive pixels, computed as:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

where TP (true positives) represents correctly segmented pixels, and FP (false positives) denotes pixels incorrectly included in the segmentation.

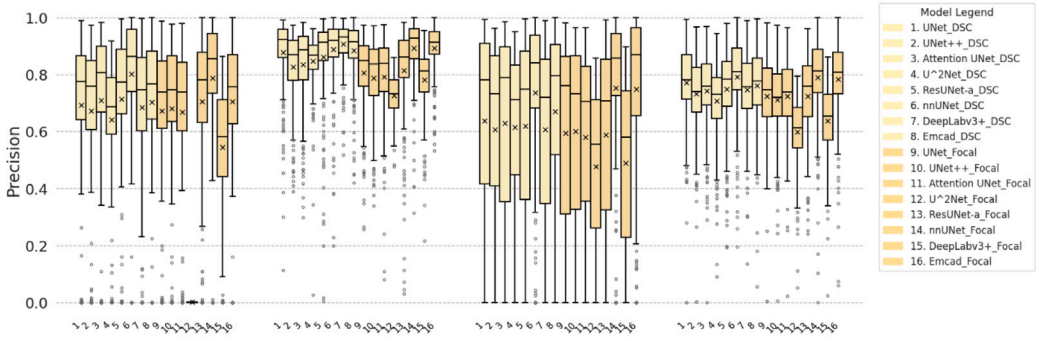
To assess the statistical significance of performance differences between Adapted SAM and CNN-based models, we employed a two-step non-parametric statistical testing procedure for each anatomical structure, using the Dice coefficient as the performance metric. First, we applied the Friedman test to determine whether at least one model differed significantly from the others across paired test samples. The test was performed independently for each anatomical class (artery, liver, stomach, and vein) using the Dice scores of all models that produced segmentations for that class. If the Friedman test returned a statistically significant result ($p < 0.05$), we conducted pairwise Wilcoxon signed-rank tests between all model pairs to identify which models exhibited statistically significant performance differences. To control for multiple comparisons, we applied the Holm–Bonferroni correction to the resulting p-values. The corrected p-value matrices were then visualized as heatmaps, where lighter colors indicate statistically significant differences between model pairs.

Table 1
Segmentation performance (median [Q3–Q1]) of CNNs and adapted SAM models in terms of (a) Intersection over Union (IoU) and (b) Dice Similarity Coefficient (DSC).

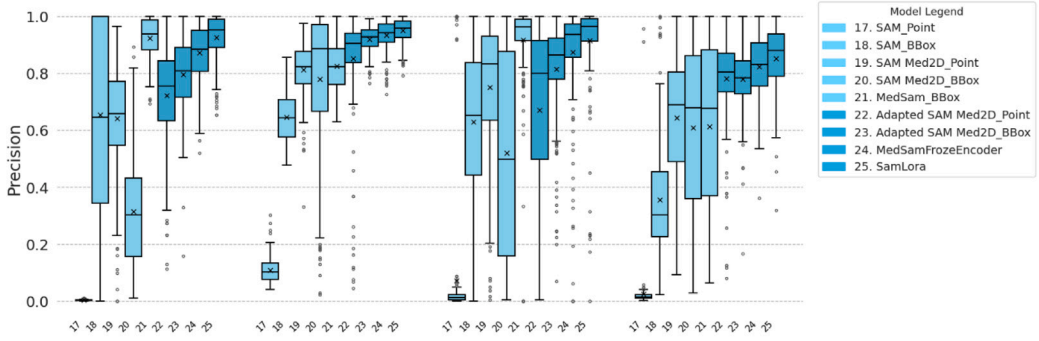
Model	Stomach	Artery	Vein	Liver
(a) IoU results (median [Q3–Q1])				
CNN Models				
$UNet_{DSC}$ (1)	0.67 [0.79 - 0.28]	0.66 [0.75 - 0.51]	0.69 [0.76 - 0.60]	0.76 [0.82 - 0.67]
$UNet++_{DSC}$ (2)	0.64 [0.78 - 0.29]	0.66 [0.74 - 0.50]	0.67 [0.73 - 0.59]	0.76 [0.80 - 0.64]
Attention $UNet_{DSC}$ (3)	0.65 [0.79 - 0.23]	0.64 [0.76 - 0.46]	0.67 [0.74 - 0.58]	0.73 [0.80 - 0.63]
U^2-Net_{DSC} (4)	0.61 [0.74 - 0.33]	0.57 [0.63 - 0.46]	0.64 [0.70 - 0.57]	0.74 [0.77 - 0.69]
ResUNet-a_{DSC} (5)	0.63 [0.77 - 0.28]	0.63 [0.73 - 0.47]	0.66 [0.71 - 0.56]	0.77 [0.82 - 0.69]
$nnUNet_{DSC}$ (6)	0.62 [0.79 - 0.03]	0.69 [0.78 - 0.53]	0.69 [0.75 - 0.60]	0.80 [0.85 - 0.73]
$DeepLabv3+_{DSC}$ (7)	0.65 [0.78 - 0.31]	0.64 [0.74 - 0.51]	0.66 [0.73 - 0.59]	0.77 [0.82 - 0.72]
$EMCAD_{DSC}$ (8)	0.62 [0.79 - 0.25]	0.66 [0.76 - 0.54]	0.67 [0.75 - 0.60]	0.80 [0.84 - 0.73]
CNN Models Trained with Focal Loss				
$UNet_{Focal}$ (9)	0.63 [0.78 - 0.21]	0.65 [0.72 - 0.52]	0.66 [0.74 - 0.58]	0.74 [0.80 - 0.65]
$UNet++_{Focal}$ (10)	0.64 [0.77 - 0.29]	0.64 [0.74 - 0.50]	0.66 [0.73 - 0.57]	0.75 [0.80 - 0.66]
Attention $UNet_{Focal}$ (11)	0.61 [0.76 - 0.28]	0.64 [0.73 - 0.49]	0.66 [0.72 - 0.58]	0.74 [0.79 - 0.63]
U^2-Net_{Focal} (12)	0.47 [0.63 - 0.23]	0.00 [0.00 - 0.00]	0.50 [0.55 - 0.44]	0.64 [0.68 - 0.60]
ResUNet-a_{Focal} (13)	0.61 [0.76 - 0.29]	0.64 [0.73 - 0.46]	0.65 [0.72 - 0.55]	0.76 [0.81 - 0.67]
$nnUNet_{Focal}$ (14)	0.65 [0.79 - 0.00]	0.69 [0.78 - 0.55]	0.69 [0.76 - 0.61]	0.80 [0.85 - 0.74]
$DeepLabv3+_{Focal}$ (15)	0.53 [0.69 - 0.20]	0.51 [0.60 - 0.38]	0.59 [0.65 - 0.52]	0.74 [0.79 - 0.67]
$EMCAD_{Focal}$ (16)	0.61 [0.78 - 0.19]	0.63 [0.71 - 0.48]	0.68 [0.74 - 0.59]	0.81 [0.85 - 0.76]
SAM Variants				
SAM_{Point} (17)	0.01 [0.02 - 0.00]	0.00 [0.00 - 0.00]	0.01 [0.02 - 0.00]	0.10 [0.07 - 0.13]
SAM_{BBox} (18)	0.42 [0.59 - 0.23]	0.22 [0.36 - 0.00]	0.26 [0.37 - 0.18]	0.60 [0.55 - 0.66]
SAM $Med2D_{Point}$ (19)	0.39 [0.66 - 0.15]	0.29 [0.41 - 0.15]	0.32 [0.43 - 0.19]	0.13 [0.08 - 0.23]
SAM $Med2D_{BBox}$ (10)	0.66 [0.79 - 0.46]	0.53 [0.63 - 0.38]	0.48 [0.59 - 0.36]	0.58 [0.67 - 0.47]
$MedSam_{BBox}$ (21)	0.64 [0.77 - 0.49]	0.63 [0.74 - 0.46]	0.47 [0.59 - 0.28]	0.70 [0.76 - 0.66]
Adapted and Fine-Tuned Variants				
Adapted SAM $Med2D_{Point}$ (22)	0.65 [0.78 - 0.37]	0.60 [0.69 - 0.48]	0.61 [0.70 - 0.48]	0.76 [0.81 - 0.67]
Adapted SAM $Med2D_{BBox}$ (23)	*0.74 [0.82 - 0.58]*	0.65 [0.73 - 0.53]	0.68 [0.73 - 0.62]	0.81 [0.84 - 0.77]
MedSamFrozeEncoder (24)	0.72 [0.81 - 0.58]	*0.70 [0.78 - 0.62]*	*0.69 [0.75 - 0.62]*	*0.82 [0.85 - 0.78]*
SamLora (25)	0.71 [0.79 - 0.59]	0.70 [0.78 - 0.59]	0.68 [0.73 - 0.59]	0.78 [0.81 - 0.74]
(b) Dice results (median [Q3–Q1])				
CNN Models				
$UNet_{DSC}$ (1)	0.80 [0.88 - 0.44]	0.80 [0.86 - 0.68]	0.82 [0.86 - 0.75]	0.86 [0.90 - 0.80]
$UNet++_{DSC}$ (2)	0.78 [0.88 - 0.45]	0.79 [0.85 - 0.66]	0.81 [0.84 - 0.74]	0.86 [0.89 - 0.78]
Attention $UNet_{DSC}$ (3)	0.79 [0.88 - 0.37]	0.78 [0.86 - 0.63]	0.80 [0.85 - 0.74]	0.85 [0.89 - 0.77]
U^2-Net_{DSC} (4)	0.76 [0.85 - 0.49]	0.73 [0.77 - 0.63]	0.78 [0.82 - 0.72]	0.85 [0.87 - 0.82]
ResUNet-a_{DSC} (5)	0.77 [0.87 - 0.44]	0.78 [0.84 - 0.64]	0.79 [0.83 - 0.72]	0.87 [0.90 - 0.82]
$nnUNet_{DSC}$ (6)	0.76 [0.88 - 0.06]	0.82 [0.88 - 0.69]	0.81 [0.86 - 0.75]	0.89 [0.92 - 0.85]
$DeepLabv3+_{DSC}$ (7)	0.79 [0.87 - 0.47]	0.78 [0.85 - 0.68]	0.80 [0.84 - 0.74]	0.87 [0.90 - 0.84]
$EMCAD_{DSC}$ (7)	0.76 [0.89 - 0.40]	0.80 [0.65 - 0.70]	0.80 [0.85 - 0.75]	0.89 [0.91 - 0.84]
CNN Models Trained with Focal Loss				
$UNet_{Focal}$ (8)	0.77 [0.88 - 0.35]	0.79 [0.84 - 0.69]	0.80 [0.85 - 0.74]	0.85 [0.89 - 0.79]
$UNet++_{Focal}$ (9)	0.78 [0.87 - 0.45]	0.78 [0.85 - 0.66]	0.80 [0.84 - 0.73]	0.86 [0.89 - 0.79]
Attention $UNet_{Focal}$ (10)	0.76 [0.87 - 0.44]	0.78 [0.84 - 0.66]	0.80 [0.84 - 0.73]	0.85 [0.88 - 0.77]
U^2-Net_{Focal} (11)	0.64 [0.77 - 0.37]	0.00 [0.00 - 0.00]	0.67 [0.71 - 0.62]	0.78 [0.81 - 0.75]
ResUNet-a_{Focal} (12)	0.76 [0.86 - 0.45]	0.78 [0.85 - 0.63]	0.79 [0.84 - 0.71]	0.86 [0.89 - 0.80]
$nnUNet_{Focal}$ (13)	0.79 [0.88 - 0.00]	0.82 [0.88 - 0.71]	0.82 [0.86 - 0.76]	0.89 [0.92 - 0.85]
$DeepLabv3+_{Focal}$ (14)	0.69 [0.82 - 0.34]	0.68 [0.75 - 0.55]	0.74 [0.79 - 0.69]	0.85 [0.88 - 0.81]
$EMCAD_{Focal}$ (7)	0.76 [0.88 - 0.31]	0.76 [0.83 - 0.64]	0.80 [0.85 - 0.74]	0.90 [0.92 - 0.86]
SAM Variants				
SAM_{Point} (15)	0.02 [0.05 - 0.01]	0.01 [0.01 - 0.00]	0.03 [0.04 - 0.02]	0.19 [0.24 - 0.14]
SAM_{BBox} (16)	0.59 [0.74 - 0.37]	0.36 [0.53 - 0.00]	0.41 [0.54 - 0.31]	0.75 [0.79 - 0.71]
SAM $Med2D_{Point}$ (17)	0.56 [0.79 - 0.27]	0.45 [0.59 - 0.26]	0.48 [0.61 - 0.32]	0.23 [0.37 - 0.15]
SAM $Med2D_{BBox}$ (18)	0.79 [0.88 - 0.63]	0.70 [0.78 - 0.55]	0.65 [0.74 - 0.53]	0.74 [0.81 - 0.64]
$MedSam_{BBox}$ (19)	0.78 [0.87 - 0.66]	0.77 [0.85 - 0.63]	0.64 [0.74 - 0.44]	0.83 [0.86 - 0.79]
Adapted and Fine-Tuned Variants				
Adapted SAM $Med2D_{Point}$ (20)	0.79 [0.88 - 0.54]	0.75 [0.82 - 0.64]	0.76 [0.82 - 0.64]	0.86 [0.90 - 0.81]
Adapted SAM $Med2D_{BBox}$ (21)	*0.85 [0.90 - 0.73]*	0.79 [0.85 - 0.69]	0.81 [0.84 - 0.76]	0.90 [0.92 - 0.87]
MedSamFrozeEncoder (22)	0.84 [0.90 - 0.73]	*0.83 [0.88 - 0.77]*	0.81 [0.84 - 0.74]	*0.90 [0.92 - 0.88]*
SamLora (23)	*0.85 [0.88 - 0.74]*	0.82 [0.87 - 0.75]	*0.82 [0.85 - 0.76]*	0.88 [0.90 - 0.86]

3. Results

Table 1 reports the median [Q3-Q1] segmentation performance of the analyzed models across the four anatomical structures (Stomach, Artery, Vein and Liver), using *DSC* and *IoU* as evaluation metrics. We compare standard CNN-based models, CNNs trained with focal loss, several variants of the SAM and adapted or fine-tuned SAM-based approaches. Table 2 compares the number of parameters involved during training. CNN models, such as nnUnet, DeepLabv3+ and EMCAD achieve moderate to high performance across organs, with the best overall results observed on the liver (e.g. $Emcad_{Focal}$: DSC 0.90 [0.92–0.86], $nnUNet_{DSC}$: DSC 0.89 [0.92–0.85]). However, their performance is notable lower on smaller or more complex structures such as the artery. Moreover, the use of focal loss does not consistently improve results and, in some cases (e.g., U^2-Net_{Focal}), leads to a drop in performance, especially for small structures like the artery and



(a) CNN Precision with Dice and Focal loss



(b) Sam Architectures Precision

Fig. 4. Boxplots showing the precision distribution of segmentation models for each anatomical class (artery, liver, stomach, vein). Panel (a) reports results from CNN-based models trained with Dice loss and Focal loss. Panel (b) presents variants of SAM, including zero-shot, pre-trained, and adapted models. The figure highlights inter-model variability and the impact of different training strategies on segmentation precision across anatomical structures.

Table 2
Comparison of model parameters.

Model	# Param(M)	# Train Param(M)
UNet	≈ 5	≈ 5
UNet++	≈ 5	≈ 5
Attention UNet	≈ 5	≈ 5
U^2-Net	≈ 4	≈ 4
ResUNet-a	≈ 6	≈ 6
DeepLabv3+	≈ 33	≈ 33
EMCAD	≈ 27	≈ 27
SAM	≈ 91	–
MedSam	≈ 91	–
Adapted SAM Med2D	≈ 271	≈ 184
MedSamFrozeEncoder	≈ 91	≈ 41
Lora-MedSam	≈ 91	≈ 41

stomach. The zero-shot variants of SAM (e.g., SAM_{Point} and SAM_{BBox}) generally exhibit poor performance, particularly on smaller and more challenging structures. For instance, SAM_{Point} achieves near-zero segmentation accuracy on both the artery and the vein. In contrast, among the SAM variants pre-trained on medical images, $SAMMed2D_{BBox}$ and $MedSAM_{BBox}$ demonstrate significantly better generalization. Among the adapted and fine-tuned variants, all models generally outperform CNN-based approaches, with the exception of Adapted SAM $Med2D_{Point}$ on the vein. The best overall performance is achieved by MedSAM-FrozenEncoder and SAM-LoRA.

Fig. 4 presents the precision distribution of the CNN-based models across the four anatomical structures. Overall, models trained with Dice loss tend to exhibit higher and more consistent precision compared to those trained with Focal loss, especially for the liver and vein, for the exception of EMCAD architecture. For the liver, almost all models achieve high precision (median values close to 0.9), with low variance. Among them, $UNet_{DSC}$, $UNet{+}{+}_{DSC}$, $ResUNet\text{-}a_{DSC}$ and $EMCAD_{Focal}$ show the most stable behavior. Precision remains high even when switching to the other focal loss experiments, though with slightly increased variability. In contrast, for the stomach, a notable drop in precision is observed for most Focal-loss models (especially $U^2\text{-}Net_{Focal}$ and $DeepLabv3^{+}_{Focal}$), with wider interquartile ranges and lower medians. The artery class is also one of the most challenging in terms of precision. While some Dice-trained models (e.g., UNetDSC, U^2-NetDSC) retain satisfactory performance, models trained with focal loss generally exhibit lower and more variable scores, with EMCAD again representing an exception. Notably, $DeepLabv3^{+}_{Focal}$ shows a sharp drop in median precision and a long tail of poor predictions. Lastly, for the vein, performance across models is relatively stable. Dice-trained networks again show slightly better precision overall, although some focal-loss variants (e.g., $UNet{+}{+}_{Focal}$, $EMCAD_{Focal}$) remain competitive. Among pre-trained variants, $SAM\ Med2D_{BBox}$ and $MedSAM_{BBox}$ show improved behavior, especially on liver and vein, where the median precision exceeds 0.7. However, the performance remains highly variable across classes, with notably poor results on artery and stomach for some models. The most consistent improvements are observed in the adapted and fine-tuned variants. In particular, SamLora and MedSamFrozeEncoder demonstrate superior and stable precision across all structures, with interquartile ranges consistently above 0.7 and limited outliers. Notably, $MedSAM_{BBox}$ achieves the highest precision on the stomach, while SamLora maintains excellent generalization across all four classes. To statistically compare the segmentation performance of all models across different anatomical structures, we computed pairwise Friedman chi-square tests based on Dice scores.

The results are visualized as heatmaps in Fig. 5, where each cell represents the statistical significance of the performance difference between two models (numbered from 1 to 25) for a given structure: artery (a), liver (b), stomach (c), and vein (d). Darker colors (values close to 1.0) indicate no statistically significant difference, while lighter colors (values closer to 10^{-4}) highlight significant performance differences. The models evaluated are as follows: $UNet_{DSC}$ (1), $UNet{+}{+}_{DSC}$ (2), Attention $UNet_{DSC}$ (3), $U^2\text{-}Net_{DSC}$ (4),

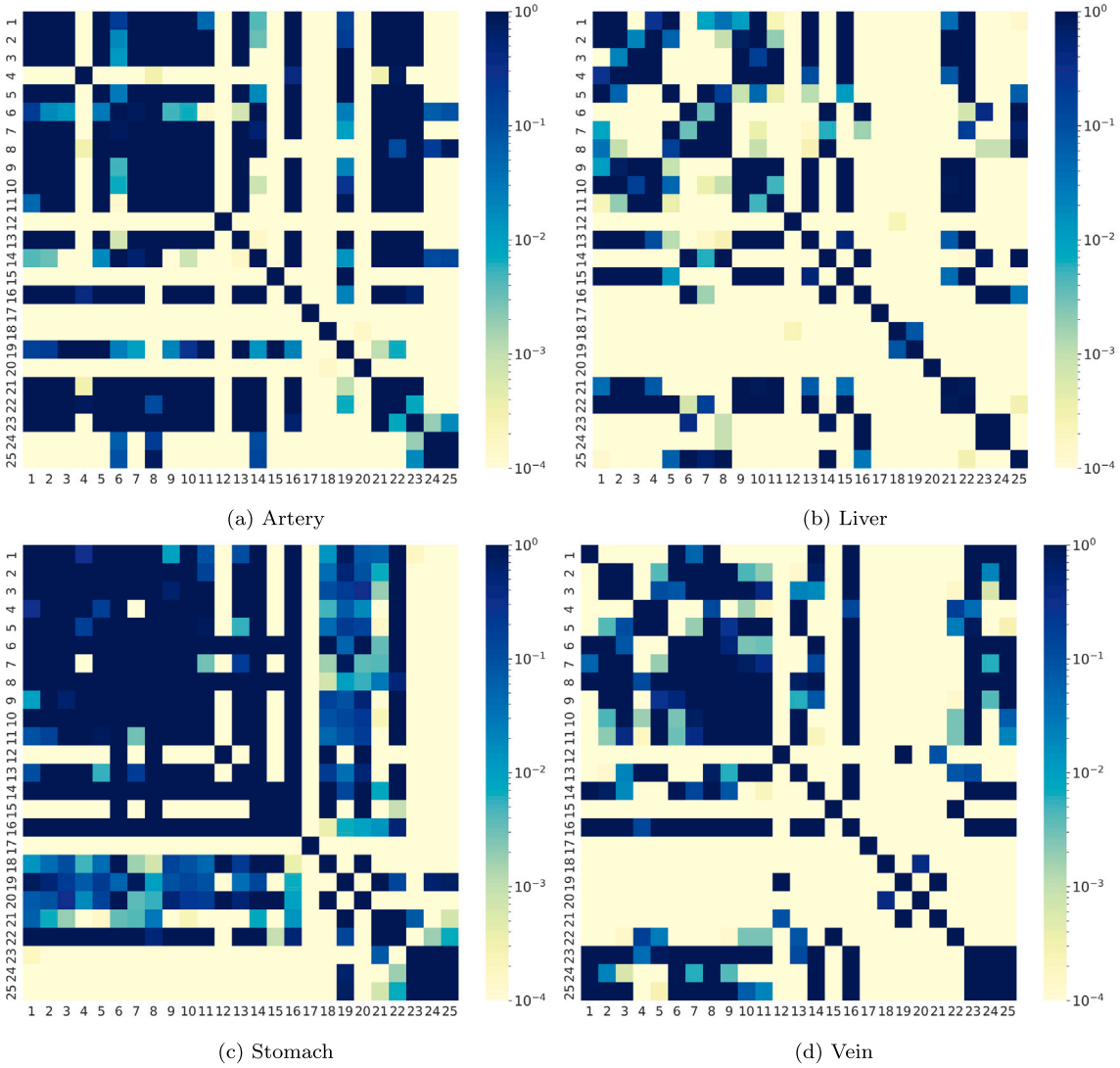


Fig. 5. Heatmaps showing pairwise statistical comparisons of segmentation performance across the 23 models for the four anatomical structures: (a) Artery, (b) Liver, (c) Stomach, and (d) Vein. The values represent Bonferroni-corrected p-values from Wilcoxon signed-rank tests, conducted only when the Friedman test indicated a significant overall difference ($p < 0.05$). Darker cells (values close to 1.0) indicate no statistically significant difference between the corresponding model pair, whereas lighter cells (values near 0.0) highlight statistically significant performance differences.

ResUNet-a_{DSC} (5), nnUNet$_{DSC}$ (6), DeepLabv3+$_{DSC}$ (7), EMCAD$_{DSC}$ (8), UNet$_{Focal}$ (9), UNet++$_{Focal}$ (10), Attention UNet$_{Focal}$ (11), U^2-Net$_{Focal}$ (12), ResUNet-a_{Focal} (13), nnUNet$_{Focal}$ (14), and DeepLabv3+$_{Focal}$ (15), EMCAD$_{Focal}$ (16), SAM$_{Point}$ (17), SAM$_{BBox}$ (18), SAM Med2D$_{Point}$ (19), SAM Med2D$_{BBox}$ (20), and MedSAM$_{BBox}$ (21), Adapted SAM Med2D$_{Point}$ (22), Adapted SAM Med2D$_{BBox}$ (23), MedSAM-FrozenEncoder (24), and SAM-LoRA (25). The heatmaps reveal several key insights regarding the performance differences between model types across anatomical structures. For small and challenging structures such as the artery and the vein, statistically significant differences between models are clearly observed, with few exceptions. In particular, zero-shot SAM variants (models 17–18) exhibit substantial performance gaps when compared to both CNN-based models and fine-tuned SAM variants, confirming their limited effectiveness in medical segmentation tasks without any form of adaptation. Moreover, adapted SAM models tend to be statistically different from CNNs, especially for the stomach which typically covers a larger portion of the image and may thus be easier to segment. This trend is also confirmed in the case of the liver, where adapted SAM models show statistically significant differences compared to most CNNs, with a few notable exceptions—most prominently nnUNet DeepLabv3+ and EMCAD, which display comparable performance. The artery and vein, on the other hand, exhibit an intermediate pattern in which CNN-based and adapted SAM models form relatively coherent performance clusters, suggesting more consistent behavior within these groups. To further assess the differences among models, Fig. 6 presents qualitative comparisons across various anatomical structures. UNet exhibits notable segmentation errors, particularly on smaller and less contrasted regions such as the artery. Although the liver is typically identified, its segmentation often lacks accuracy in both shape and spatial alignment. nnUNet provides more reliable results, with improved delineation of both large and small structures; nonetheless, occasional oversegmentation and geometric distortions are still evident. MedSAM-FrozenEncoder produces segmentations that closely match the ground truth, especially for larger organs like the liver and stomach, and demonstrates consistent performance in identifying smaller vascular structures. Lastly, SAM-LoRA also achieves strong results, striking a balance between accurate segmentation of large organs and effective detection of finer anatomical details.

4. Discussions

The accurate segmentation of fetal abdominal structures in US imaging is essential for assessing fetal development, detecting anomalies,

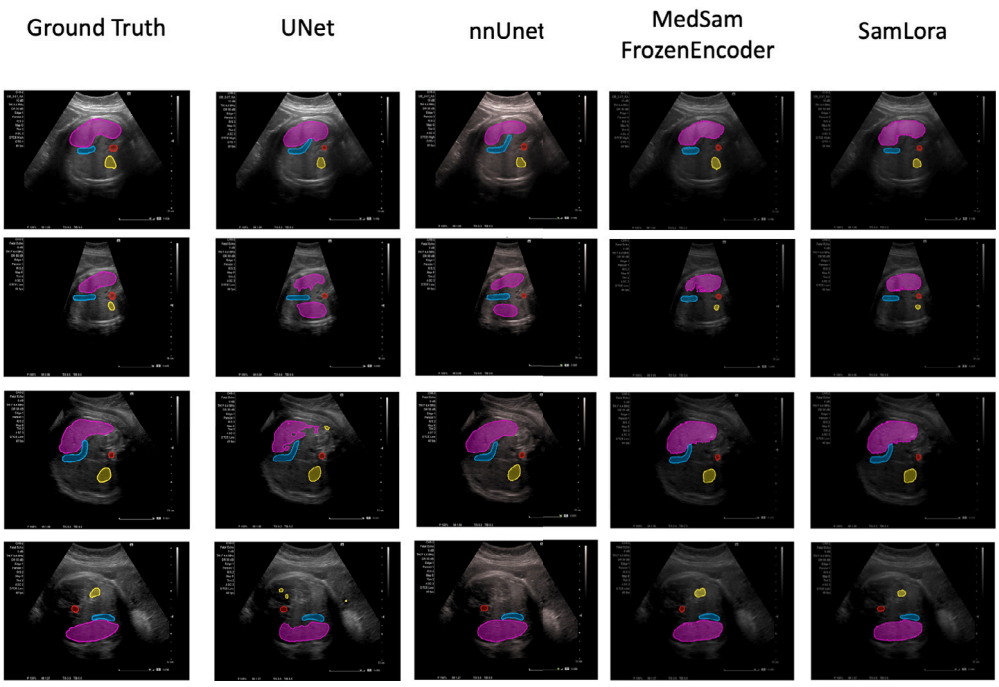


Fig. 6. Qualitative segmentation results on fetal abdominal ultrasound images using different models. The figure showcases representative examples from the test set, comparing predictions from UNet, nnUNet, MedSAM-FrozenEncoder, and SAM-LoRA. Each image includes ground truth annotations alongside the predicted masks, allowing visual assessment of segmentation quality across various anatomical structures: artery, liver, stomach, and vein.

and supporting clinical decision-making during prenatal care. Unlike the fetal head, which has been extensively studied in the literature, the fetal abdomen presents a more complex and variable anatomy, with overlapping organs, heterogeneous echogenicity, and frequent artifacts. Structures such as the liver, stomach, arteries, and veins offer critical insights into fetal well-being but are particularly challenging to delineate manually due to their subtle boundaries and visual similarity. This complexity reinforces the clinical need for reliable and efficient automated segmentation methods, which can improve consistency, reduce workload, and potentially support large-scale screening programs. In this context, our study addresses a fundamental question: what is the most effective strategy for segmenting internal fetal abdominal structures—adapting a general-purpose foundation model like SAM, or relying on specialized convolutional architectures tailored to the task? Which approach offers the best trade-off between accuracy and clinical applicability?

CNN-based models — such as UNet, nnUNet, DeepLabv3+ and EMCAD — achieve moderate to high segmentation performance, particularly on well-defined and relatively large anatomical structures like the liver. Notably, $\text{nnUNet}_{\text{DSC}}$ achieved a median Dice score of 0.89[0.92–0.85] on the liver, despite being trained from scratch without pretrained weights. This highlights the effectiveness of task-specific architectures in capturing relevant anatomical features, even in data-constrained scenarios, as long as organ boundaries are sufficiently distinct. However, these same models struggled with smaller, less well-defined — such as arteries and veins — where performance dropped significantly. This degradation can be attributed to several factors: the small spatial footprint of these structures makes them more susceptible to being overlooked by convolutional filters, their low contrast and inconsistent appearance across samples reduce the model's ability to learn robust features, and class imbalance further biases predictions towards larger, more prominent regions. As a result, CNNs tend to under-segment or completely miss these challenging structures, particularly when trained on limited and heterogeneous US data. Notably, the use of focal loss did not consistently enhance segmentation performance, particularly for small anatomical structures, where a noticeable drop in Dice and precision was frequently observed. Although focal loss was originally proposed to mitigate class imbalance by emphasizing hard-to-classify examples, its effectiveness appears limited in the context of fetal US. This may be due to the inherent noise, low contrast, and variable appearance typical of US imaging, which can cause the model to misinterpret meaningful anatomical boundaries as hard negatives. As a result, focal loss may over-penalize ambiguous but correct predictions, leading to suboptimal segmentation. This issue is especially pronounced in architectures such as $\text{U}^2\text{-Net}_{\text{Focal}}$ and $\text{DeepLabv3+}_{\text{Focal}}$, which exhibit severe performance degradation.

A similar trend is observed for the stomach, where despite being a relatively larger structure, a noticeable drop in performance — particularly in terms of precision — is evident. As shown in Fig. 3, this can be attributed to the high variability in the stomach's appearance and shape across acquisition conditions. Such variability challenges the model's ability to consistently delineate its boundaries, resulting in frequent over- or under-segmentation.

As regards zero-shot SAM variants (e.g., SAMPoint and SAMBBox), they perform poorly across all structures, particularly on the artery and vein, where near-zero Dice scores were observed. These findings reinforce the conclusion that SAM, in its original configuration, is not suitable for direct deployment in fetal US segmentation tasks—as also demonstrated in previous work on other medical imaging tasks [39]. However, the performance improved substantially with SAM variants pre-trained on medical images, such as MedSAMBBox and SAM Med2DBBox, which exhibited better generalization, particularly for liver and vein segmentation. Nonetheless, our results demonstrate that even when using models such as MedSAM — already trained on fetal US data such as HC18 [40] — further adaptation remains necessary to achieve reliable segmentation across all structures. This suggests that the challenge lies not in a domain shift from natural to medical images, but in the intrinsic variability within fetal US itself. Factors such as low and inconsistent echogenicity, frequent overlaps between adjacent structures, and operator-dependent acquisition angles contribute to significant intra-class variability. These issues are especially problematic for small and poorly contrasted organs like the artery and stomach, where precise localization is critical. For these reasons, lightweight

adaptation techniques such as LoRA, Adaptive Layers or encoder freezing are essential to tailor the model's internal representations to the fine-grained appearance patterns specific to each structure, thereby improving both segmentation accuracy and consistency. The most consistent improvements are in fact observed with the adapted and fine-tuned version of SAM, with MedSamFrozeEncoder and SamLora achieving the highest DSC (similar to [41]). In contrast, Adapted SAM $\text{Med2D}_{\text{BBox}}$, while still yielding strong results, showed slightly reduced performance and higher variability on smaller structures such as the artery and stomach. This performance gap can be attributed to multiple factors. First, SamLoRA modifies a small set of low-rank matrices within the transformer layers, enabling targeted adaptation with a minimal increase in trainable parameters. This not only improves generalization, but also reduces the risk of overfitting—especially relevant in fetal US, where annotated data is limited and intra-class variability is high. In contrast, Adapted SAM $\text{Med2D}_{\text{BBox}}$ involves adapting a larger portion of the network. While this may increase expressive power, it also significantly increases the number of trainable parameters (Table 2), which may lead to instability during training, especially in low-data regimes. Similarly, MedSAM-FrozenEncoder, which retains the pre-trained encoder in a frozen state while fine-tuning only the segmentation head, demonstrated stable performance. This suggests that adapting solely the segmentation head may be sufficient to achieve robust results, without the need to update the entire network. However, a slight drop in precision was observed compared to LoRA-based fine-tuning, as also confirmed by the qualitative results shown in Fig. 6. This degradation may be attributed to the limited adaptability of the frozen encoder, which might not fully capture subtle domain-specific features or fine-grained anatomical variations present in the target data.

Beyond segmentation accuracy, the clinical applicability of adapted foundation models like SAM hinges on their interactivity, interpretability, and integration within real-world workflows [42]. In this context, prompt-based models offer a significant advantage: their design naturally supports human-in-the-loop interaction, enabling clinicians or sonographers to guide or refine segmentations with minimal effort. Among the different prompting strategies evaluated, our results highlight that bounding box prompts consistently outperform point-based prompts. This is expected, as point-based prompts provide limited spatial information and are more sensitive to both the number and precise placement of the points. Even small inaccuracies in point location can significantly affect the segmentation outcome, especially in fetal abdominal US where anatomical structures may be small, partially visible, or affected by imaging artifacts. In contrast, bounding boxes offer a more robust and informative spatial prior, helping the model better localize the target region and reducing the impact of noisy inputs. Moreover, this functionality could be highly valuable since a sonographer could quickly draw a bounding box around a suspected structure, triggering a high-quality segmentation without the need for full manual delineation. This interaction paradigm not only reduces annotation time, but also enhances trust and usability, as the user retains control over the process. Moreover, this interactive ability could serve as a training tool, enabling less experienced clinicians to visualize and learn from model-generated segmentations, adjusted in real-time. However, these interactive advantages must be weighed against practical deployment considerations. As shown in Table 2, SAM-based models (SAM-Med2D and MedSAM) contain approximately 93.7M parameters, representing roughly a 3-fold increase compared to nnU-Net'ones. This substantially larger model size directly impacts memory footprint and inference speed, which are critical factors for real-time clinical deployment, particularly on resource-constrained devices commonly used in point-of-care US settings. Therefore, the choice between task-specific CNNs and adapted foundation models should consider not only segmentation accuracy and interactive capabilities, but also the computational resources available in the target clinical environment. While our study provides a first step towards model segmentation for fetal abdominal structures in US, several limitations must be acknowledged. First and foremost, the size of the dataset used for training and evaluation remains limited. The relatively small number of samples may restrict the model's ability to generalize to the full spectrum of fetal anatomical variability observed in real-world clinical practice. This limitation is particularly relevant for the smallest and least contrasted structures, such as the artery and stomach, where segmentation performance was more variable even among the best-performing models. Although standard data augmentation techniques were applied to increase variability during training, their impact is inherently limited by the original dataset size and diversity [43]. More advanced augmentation strategies, including the use of generative models to synthesize realistic anatomical variations, could further enhance model robustness, particularly for small and underrepresented structures, drawing inspiration from recent works in the field [44–46]. Second, while we explored a range of adaptation strategies for SAM, hyperparameter tuning and layer selection were not exhaustively optimized. Future work should investigate more granular adaptation polices, potentially integrating task-specific prompt tuning or adapter-based learning at different transformer depths to balance expressiveness and regularization. In addition, given the growing interest in Transformer-based architectures for medical image segmentation, future research could also explore their integration within this task, either as standalone models or in hybrid approaches, to assess their potential advantages over CNNs and SAM-based methods in fetal abdominal segmentation. Furthermore, while our study focused on comparing general-purpose foundation models with task-specific CNNs to address the "adapt or specialize" question, a comprehensive comparison with US-specific foundation models such as USFM [47], URFM [48], represents a valuable direction for future research. Such an investigation could further elucidate the trade-offs between general-purpose, medical-specific, and US-specific foundation models in the context of fetal US analysis. Third, our evaluation relied on 2D static frames extracted from fetal US scans. However, in clinical scenarios, temporal consistency across video sequences is often important, especially when assessing dynamic structures. Recently, SAM2 [49] and its medical variant, MedicalSAM2 [50], have been proposed, introducing significant improvements in handling video sequences and enabling real-time segmentation through memory mechanisms and efficient mask propagation. Although our work focuses on frame-based 2D segmentation, as the task is limited to static segmentation, integrating SAM2's video capabilities represents a promising future direction to better exploit the temporal continuity inherent in US acquisition. Additionally, hybrid architectures such as [51] could be explored, combining SAM's robust feature extraction with UNet's proven ability to preserve fine-grained spatial details through skip connections.

5. Conclusions

This study provides the first comprehensive evaluation of segmentation models for the segmentation of internal fetal abdominal structures in US. By comparing multiple adapted variants of SAM with well-established, task-specific CNNs, we demonstrate that while traditional CNNs remain competitive for segmenting large, high-contrast organs, they struggle with smaller or poorly defined structures such as the artery and stomach. While zero-shot SAM variants are not suitable for fetal abdominal segmentation, adapted versions of SAM achieve competitive or superior performance compared to traditional CNN-based models. Fine-tuning proves essential to address the specific challenges of fetal US, such as anatomical variability and low contrast.

In addition to their segmentation accuracy, SAM-based models offer clinical advantages through prompt-based interaction, enabling efficient human-in-the-loop workflows. These findings suggest that adapted foundation models represent a promising and flexible solution for automated analysis in fetal ultrasound imaging.

CRediT authorship contribution statement

Maria Chiara Fiorentino: Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lorenzo Federici:** Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis, Data curation, Conceptualization. **Alessandro Pietro La Camera:** Visualization, Validation, Software, Investigation, Data curation. **Enrico Gianluca Caiani:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Conceptualization.

Funding

The authors declare that for this study no benefits have been or will be received.

Ethical approval

No ethics approval was required.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.C. Fiorentino, F.P. Villani, M. Di Cosmo, E. Frontoni, S. Moccia, A review on deep-learning algorithms for fetal ultrasound-image analysis, Med. Image Anal. 83 (2023) 102629.
[2] S.K.B. Degala, R.P. Tewari, P. Kamra, U. Kasiviswanathan, R. Pandey, Segmentation and estimation of fetal biometric parameters using an attention gate double u-net with guided decoder architecture, Comput. Biol. Med. 180 (2024) 109000.
[3] L. Salomon, Z. Alfirevic, V. Berghella, C. Bilardo, G. Chalouhi, F.D.S. Costa, E. Hernandez-Andrade, G. Malinger, H. Munoz, D. Paladini, et al., Isuog practice guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan, Ultrasound. Obstet. Gynecol 59 (2022) 840–856.
[4] N. Hernandez-Cruz, O. Patey, C. Teng, A.T. Papageorghiou, J.A. Noble, A comprehensive scoping review on machine learning-based fetal echocardiography analysis, Comput. Biol. Med. 186 (2025) 109666.
[5] G. Chen, J. Bai, Z. Ou, Y. Lu, H. Wang, Psfhs: intrapartum ultrasound image dataset for ai-based segmentation of pubic symphysis and fetal head, Sci. Data 11 (436) (2024).
[6] L. Liu, D. Tang, X. Li, Y. Ouyang, Automatic fetal ultrasound image segmentation of first trimester for measuring biometric parameters based on deep learning, Multimedia Tools Appl. 83 (2024) 27283–27304.
[7] D. Shen, G. Wu, H.I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.
[8] H.P. Chan, R.K. Samala, L.M. Hadjiiski, C. Zhou, Deep learning in medical image analysis, in: Deep Learning in Medical Image Analysis: Challenges and Applications, 2020, pp. 3–21.
[9] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S.X. Li, D. Ni, T. Wang, Deep learning in medical ultrasound analysis: a review, Engineering 5 (2019) 261–275.
[10] S. Moccia, M.C. Fiorentino, E. Frontoni, Mask-r 2 cnn: a distance-field regression version of mask-rcnn for fetal-head delineation in ultrasound images, Int. J. Comput. Assist. Radiol. Surg. 16 (2021) 1711–1718.
[11] A.A. Hekal, H.M. Amer, H.E.D. Moustafa, A. Elnakib, Automatic measurement of head circumference in fetal ultrasound images using a squeeze atrous pooling unet, Biomed. Signal Process. Control. 103 (2025) 107434.
[12] S. Parvathavarthini, K. Sharvanthika, S. Sindhu, K. Kaviya, Fetal head circumference measurement from ultrasound images using attention u-net, in: 2023 Fifth International Conference on Electrical, Computer and Communication Technologies, ICECCT, IEEE, 2023, pp. 1–5.
[13] H. Verma, B. Patro, Enhanced fetal ultrasound image segmentation using spatial attention mechanisms with unet: Saunet, in: 2024 IEEE International Conference on Computer Vision and Machine Intelligence, CVMI, IEEE, 2024, pp. 1–6.
[14] L. Halder, S. Mohanto, M. Islam, M.T. Jawad, Fetal head segmentation and head circumference measurement using residual u-net, in: 2024 2nd International Conference on Information and Communication Technology, ICICT, IEEE, 2024, pp. 145–149.
[15] S.M. Amini, Head circumference measurement with deep learning approach based on multi-scale ultrasound images, Multimedia Tools Appl. 81 (2022) 32981–32993.
[16] B. Kim, K.C. Kim, Y. Park, J.Y. Kwon, J. Jang, J.K. Seo, Machine-learning-based automatic identification of fetal abdominal circumference from ultrasound images, Physiol. Meas. 39 (2018) 105007.
[17] S. Zhang, D. Metaxas, On the challenges and perspectives of foundation models for medical image analysis, Med. Image Anal. 91 (2024) 102996.
[18] J. Ambsdorf, A. Munk, S. Llambias, A.N. Christensen, K. Mikolaj, R. Balestriero, M.G. Tolsgaard, A. Feragen, M. Nielsen, General methods make great domain-specific foundation models: A case-study on fetal ultrasound, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2025, pp. 271–281.
[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
[20] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017.
[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[22] D. Alexey, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[24] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, R. Ng, Fourier features let networks learn high frequency functions in low dimensional domains, Adv. Neural Inf. Process. Syst. 33 (2020) 7537–7547.
[25] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, et al., Sam-med2d, 2023, arXiv preprint arXiv:2308.16184.
[26] K. Zhang, D. Liu, Customized segment anything model for medical image segmentation, 2023, arXiv preprint arXiv:2304.13785.
[27] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models, in: ICLR, vol. 1, 2022, p. 3.
[28] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat. Commun. 15 (654) (2024).
[29] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021, arXiv preprint arXiv:2106.09685.
[30] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.
[31] Z. Zhou, M.M. Rahma. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.
[32] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, 2018, arXiv preprint arXiv:1804.03999.
[33] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O.R. Zaiane, M. Jagersand, U2-net: Going deeper with nested u-structure for salient object detection, Pattern Recognit. 106 (2020) 107404.
[34] F.I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data, ISPRS J. Photogramm. Remote Sens. 162 (2020) 94–114.
[35] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2021) 203–211.
[36] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.
[37] M.M. Rahman, M. Munir, R. Marculescu, Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11769–11779.
[38] C.H. Kusters, T.J. Jaspers, T.G. Boers, M.R. Jong, J.B. Jukema, K.N. Fockens, A.J. d. Groof, J.J. Bergman, F.van der. Sommen, P.H. D. With, Will transformers change gastrointestinal endoscopic image analysis? a comparative analysis between cnns and transformers, in terms of performance, robustness and generalization, Med. Image Anal. 99 (2025) 103348.
[39] X. Li, Q. Hu, X. Lin, Y. Li, Y. Dong, T. Lin, Echosam: Sam adaption for unified 2d echocardiography segmentation and ejection fraction calculation, Biomed. Signal Process. Control. 109 (2025) 108000.

[40] T.L. van den Heuvel, D. de Bruijn, C.L. de Korte, B.V. Ginneken, Automated measurement of fetal head circumference using 2d ultrasound images, PloS One 13 (2018) e0200412.

[41] X. Chen, C. Wang, H. Ning, S. Li, M. Shen, Sam-octa: Prompting segment-anything for octa image segmentation, Biomed. Signal Process. Control. 106 (2025) 107698.

[42] Y. Zhang, Z. Shen, R. Jiao, Segment anything model for medical image segmentation: Current applications and future directions, Comput. Biol. Med. (2024) 108238.

[43] M.C. Fiorentino, S. Moccia, M.D. Cosmo, E. Frontoni, B. Giovanola, S. Tiribelli, Uncovering ethical biases in publicly available fetal ultrasound datasets, Npj Digit. Med. 8 (355) (2025).

[44] Y. Duan, T. Tan, Z. Zhu, Y. Huang, Y. Zhang, R. Gao, P.C.I. Pang, X. Gao, G. Tao, X. Cong, et al., Fetalflex: Anatomy-guided diffusion model for flexible control on fetal ultrasound image synthesis, 2025, arXiv preprint arXiv:2503.14906.

[45] F. Wang, K. Whelan, G. Silvestre, K.M. Curran, Generative diffusion model bootstraps zero-shot classification of fetal ultrasound images in underrepresented african populations, in: International Workshop on Preterm, Perinatal and Paediatric Image Analysis, Springer, 2024, pp. 143–154.

[46] A. Lasala, M.C. Fiorentino, S. Micera, A. Bandini, S. Moccia, Exploiting class activation mappings as prior to generate fetal brain ultrasound images with gans, in: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2023, pp. 1–4.

[47] J. Jiao, J. Zhou, X. Li, M. Xia, Y. Huang, L. Huang, N. Wang, X. Zhang, S. Zhou, Y. Wang, et al., Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis, Med. Image Anal. 96 (2024) 103202.

[48] Q. Kang, Q. Lao, J. Gao, W. Bao, Z. He, C. Du, Q. Lu, K. Li, Urfm: A general ultrasound representation foundation model for advancing ultrasound image diagnosis, IScience 28 (2025).

[49] N. Ravi, V. Gabeur, Y.T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al., Sam 2: Segment anything in images and videos, 2024, arXiv preprint arXiv:2408.00714.

[50] J. Zhu, A. Hamdi, Y. Qi, Y. Jin, J. Wu, Medical sam 2: Segment medical images as video via segment anything model 2, 2024, arXiv preprint arXiv:2408.00874.

[51] X. Xiong, Z. Wu, S. Tan, W. Li, F. Tang, Y. Chen, S. Li, J. Ma, G. Li, Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation, 2024, arXiv preprint arXiv:2408.08870.