A New Multicenter Testicular US Dataset and a Lightweight Cond-UNet for Generalization in US Segmentation

Nicola Morelli[1,*]
nicola.morelli@unimore.it
Kevin Marchesini[1,*]
kevin.marchesini@unimore.it
Daniele Santi[2]
daniele.santi@unimore.it
Costantino Grana[1]
costantino.grana@unimore.it
Federico Bolelli[1,†]
federico.bolelli@unimore.it

[1] Department of Engineering "Enzo Ferrari" University of Modena and Reggio Emilia, Italy

[2] Department of Biomedical, Metabolic and Neural Sciences University of Modena and Reggio Emilia, Italy

[*] Equal contribution.

[†] Corresponding author.

Abstract

Male infertility is a significant yet under-addressed global health condition, and testicular ultrasound (US) plays a central role in its diagnostic evaluation. We introduce and publicly release TesticulUS-Real, the first multicenter testicular US segmentation dataset, addressing the absence of annotated public benchmarks for this anatomy. The dataset comprises 1,053 real ultrasound images acquired from two independent clinical institutions, with expert segmentation masks obtained through a standardized annotation and consensus review protocol. Leveraging this resource, we define an open-organ segmentation protocol to evaluate how models trained on existing multi-organ US data transfer to a previously unseen anatomical target. Beyond the dataset release, we conduct a broad cross-organ segmentation generalization study on ultrasound data. Using the UUSIC benchmark, we evaluate generalization across five anatomical regions and independent acquisition centers, comparing task-specific segmentation models, generalization-oriented ultrasound methods, and SAM-based foundation models under fully automatic inference. Alongside this benchmark, we introduce Cond-UNet, a lightweight conditional U-Net that combines Feature-wise Linear Modulation (FiLM) with our newly proposed shared attention conditioning (SAC) to obtain adaptive organ-aware representations. Experiments show that Cond-UNet achieves the best average cross-organ generalization performance across the UUSIC organs while using fewer parameters and lower computational cost than foundation-model alternatives. In the open-organ setting, the proposed testicular dataset enables a direct analysis of how different model families behave when facing an unseen ultrasound anatomy, highlighting the role of large-scale pretraining for foundation models and the robustness of organ-aware conditioning in lightweight architectures. The code is publicly released at https://github.com/AImageLab-zip/US_Cond-UNet, and the dataset at https://ditto.ing.unimore.it/testiculus/.

1 Introduction

Ultrasound (US) imaging is one of the most widely used diagnostic modalities due to its non-invasive, real-time capability, low cost, portability, and safety profile. These properties make it particularly important in anatomical regions where repeated, accessible, and radiation-free imaging is desirable. At the same time, despite its widespread adoption, automatic ultrasound segmentation remains highly challenging. Compared with many other imaging modalities, US images are strongly affected by speckle noise, a low signal-to-noise ratio, operator dependency, anatomical variability, and substantial domain shifts across scanners and institutions, which significantly affect model robustness [15]. These factors make generalization across centers a central open problem in ultrasound image analysis [15, 41, 45].

One clinically important but underrepresented application area is testicular ultrasound (testicular US or TUS). Infertility affects approximately one in six adults worldwide, according to the World Health Organization, highlighting the need for accessible diagnostic tools and improved reproductive health assessment pathways [35]. Testicular US plays an important role in the evaluation of male reproductive health, including the assessment of testicular morphology, volume, focal abnormalities, and other imaging findings relevant to clinical work-up [26, 41]. Accurate testicular delineation can support a more objective and precise analysis by enabling automated measurements of organ size and shape, quantitative assessment of parenchymal echotexture, and localization of focal or diffuse abnormalities within the testicular region. Despite this clinical relevance, public datasets for testicular US segmentation remain absent. This lack of data hinders the development, comparison, and reproducible evaluation of automatic segmentation methods for this anatomy.

To the best of our knowledge, the only existing public resource related to testicular ultrasound is our previously introduced TesticulUS [33] dataset, which releases deep-learning-generated synthetic images without providing any type of label. In this work, we introduce and publicly release a *new multicenter testicular ultrasound segmentation dataset*, named TesticulUS-Real. The dataset contains 1,053 real US images with corresponding testicular segmentation masks, acquired from two independent clinical institutions and annotated by clinical experts using a standardized protocol. Our dataset therefore fills an important gap by providing a public real TUS dataset to enable, for the first time, testicular ultrasound segmentation training for deep learning algorithms.

In the last decade, deep learning has become the dominant paradigm for automatic ultrasound segmentation, with U-Net-based architectures [16, 17, 39] providing competitive fully automatic baselines. More recently, foundation models, such as SAM [19], have been explored for their potential to improve generalization across datasets and anatomical targets. In the medical domain a model that generalizes across organs and centers would simplify deployment; however, many SAM-based methods [9, 29, 47] require informative prompts, such as bounding boxes, clicks, or masks, to specify the structure of interest to work effectively. In practice, such prompts may require expert input or be derived from annotation-like information, establishing a strong dependency on external supervision at inference time, limiting full automation and introducing an implicit form of information leakage. In realistic clinical scenarios where no annotation guidance is available, this reliance can substantially compromise robustness and fairness in evaluation [1, 20, 34, 46]. This motivates our study of cross-organ and cross-center ultrasound segmentation under fully automatic inference.

Alongside the dataset, we propose a *lightweight conditional segmentation network* for cross-organ ultrasound segmentation, designed to generalize across different organs and be robust to data acquired from different clinical centers. The model is based on a U-Net back-

bone modulated through Feature-wise Linear Modulation (FiLM) [37] layers, combined with our proposed shared attention conditioning (SAC) module that enables organ-aware adaptation within a compact architecture.

We systematically benchmark our method against state-of-the-art segmentation models and vision foundation models on five different organs and on data coming from independent centers. For this purpose, we use the UUSIC [24] dataset, introduced in the homonymous challenge, which provides a segmentation multicenter ultrasound benchmark spanning five anatomical regions, including breast, cardiac, fetal head, kidney, and thyroid. This setting enables the evaluation of robustness under domain shifts across independent institutions.

Finally, we use our newly introduced testicular US dataset to define an open-organ evaluation protocol, where models trained on existing multi-organ ultrasound data are tested on a previously unseen anatomical target. This evaluation is enabled only by our dataset, because no public real testicular ultrasound segmentation dataset existed before, making testicular US an anatomical domain unseen by pretrained segmentation foundation models. Notably, in the cross-organ task, our lightweight architecture achieves superior performance without relying on large-scale pretraining or inference-time supervision.

Contributions. In summary, the contributions of this work are threefold: (*i*) we publicly release a novel multicenter testicular ultrasound segmentation dataset, providing real images with expert testicular masks and addressing a missing anatomical region in public US segmentation benchmarks; (*ii*) we introduce a lightweight conditional segmentation network for cross-organ US segmentation, based on a U-Net backbone modulated with FiLM layers and a shared attention conditioning module, as a compact baseline able to generalize across organs and clinical centers; (*iii*) we provide a systematic benchmark against state-of-the-art task-specific segmentation models, and SAM-based foundation models under cross-organ generalization and open-organ evaluation, using UUSIC for multi-organ multicenter testing and our testicular dataset as an unseen anatomical target.

2 Related Work

Medical image segmentation has traditionally been dominated by task-specific supervised models [4, 48], where the architecture, training protocol, and supervision are optimized for a fixed anatomy, modality, or annotation setting. More recently, large segmentation foundation models, especially SAM-style promptable models, have shifted attention toward approaches that can generalize across datasets and tasks [3, 31]. However, models pretrained on natural images often require medical-domain adaptation, fine-tuning, or adapters to work reliably on medical images, where appearance, boundaries, and patterns differ substantially [1, 55]. This challenge is even stronger in ultrasound segmentation, where speckle noise, low contrast, artifacts, and weak boundaries make robust delineation particularly difficult [15]. As a result, task-specific models remain strong baselines, while recent works have started adapting foundation models to ultrasound through fine-tuning or automatic prompting.

2.1 Task-Specific Segmentation Models

We refer to task-specific segmentation models as architectures and training pipelines optimized for a particular segmentation problem, usually defined by a target anatomy, imaging modality, and annotated dataset. U-Net [39] was the first important model for medical image

segmentation, establishing the encoder-decoder architecture with skip connections as a standard paradigm for this task. Building on the U-Net architecture, nnU-Net [16, 17] became a strong reference baseline by automatically configuring preprocessing, architecture hyperparameters, training, and post-processing for each target task. Several medical segmentation variants of U-Net have also been proposed, introducing residual [2] or densely connected blocks [51] to improve feature extraction and training stability. Nevertheless, nnU-Net, particularly with residual encoders [17], remains one of the best-performing methods across many medical image segmentation benchmarks.

Although CNN-based models achieve strong performance, their local receptive fields can limit their ability to capture long-range dependencies and global anatomical context. To address this limitation, several works have integrated Transformer self-attention into U-Net-like architectures [18], improving contextual modeling while preserving the localization benefits of encoder-decoder designs. Representative methods in this direction include TransUNet [7], UNETR [13], Swin-UNETR [12], and nnFormer [58].

More recently, state-space models, particularly Mamba-based architectures [11], have emerged as efficient alternatives to Transformers for long-range dependency modeling. Originally developed for sequence modeling, Mamba has been rapidly extended to 2D and 3D vision tasks [25], including a growing number of applications in medical image analysis [21]. One of the first works in this direction was U-Mamba [30], which combines convolutional feature extraction with state-space modeling and retains a self-configuring strategy for biomedical segmentation. Related methods, such as SegMamba [49], LKM-UNet [44], MultiSegMamba [27, 28], and IM-Fuse [38], further explore Mamba-based designs trying to improve local and global spatial modeling, also exploring multi-directional feature scanning.

In ultrasound segmentation, task-specific models must also cope with limited and noisy annotations. For this reason, semi-supervised methods have been explored, such as Shape-Prior-Semi-Seg (SPSS) [8], which introduces a simple prior-aware pseudo-labeling strategy with an adversarially learned shape prior. Only a few works explicitly target generalization across unseen ultrasound domains [6, 53, 54]. Among them, the most relevant is MI-SegNet [6], which disentangles anatomical and domain-specific representations, encouraging segmentation to rely on domain-independent anatomical features. These works are especially relevant to our setting, where robustness must be evaluated across ultrasound organs and centers, rather than only under in-distribution conditions.

2.2 Segmentation Foundation Models

In parallel with task-specific segmentation models, recent work has explored segmentation foundation models, largely inspired by the Segment Anything Model (SAM) [19]. SAM introduced a promptable segmentation framework composed of an image encoder, a flexible prompt encoder, and a mask decoder, enabling segmentation from points, boxes, or masks. However, direct transfer from natural images to medical images is limited by the domain gap, motivating many medical SAM adaptations. These include adapter-based methods such as Med-SA [47], large-scale medical fine-tuning as in MedSAM [29], and general-purpose 2D/3D medical variants such as SAM-Med2D [9] and SAM-Med3D [43]. MedSAM is particularly relevant as a generalist medical segmentation foundation model trained to support segmentation across multiple modalities and anatomical structures.

Ultrasound-specific SAM adaptations have also been proposed to address the domain gap between natural images and ultrasound. SAMUS [23] adapts SAM to ultrasound by combining the original image encoder with a CNN branch for local features and lightweight

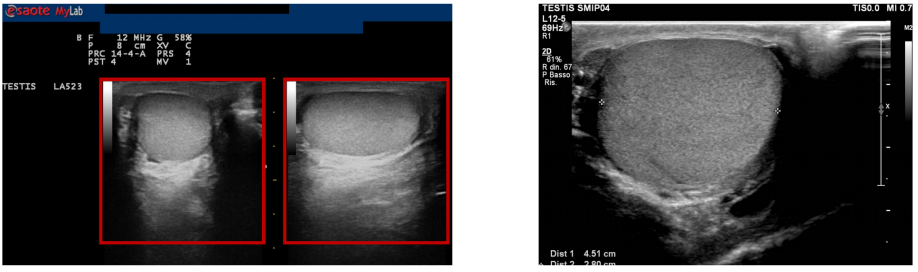

Figure 1: Sample images from the Esaote and Philips machines. The red boxes indicate the cropped regions retained after preprocessing.

feature/position adapters. UltraSAM [32] follows a data-centric strategy by assembling a large collection of public ultrasound segmentation datasets and training a SAM-style ultrasound foundation model under a prompt-conditioned segmentation paradigm.

Despite their generalization potential, SAM-style models often rely on prompts, i.e., points, boxes, or masks, which may need expert interaction at inference time. To reduce this dependency, recent works have explored automatic prompting. In ultrasound, AutoSAMUS [23] extends SAMUS with an auto-prompt generator for end-to-end segmentation, while APG-SAM [50] learns automatic prompts for breast US lesion segmentation.

Overall, these two lines of work reveal a central trade-off in ultrasound segmentation: task-specific models achieve good performance, are efficient, and fully automatic, while foundation models offer broader generalization but remain limited by prompt dependency and cross-organ robustness. Our work aims to bridge this gap with a *lightweight task-specific-style architecture* designed to *generalize across ultrasound organs and pathologies*.

3 Methods

3.1 TesticulUS-Real: A Multicenter Testicular Ultrasound Dataset

In this work, we introduce TesticulUS-Real, a new multicenter testicular ultrasound segmentation dataset containing 1,053 real B-mode ultrasound images and expert segmentation masks. Data were collected from two independent clinical institutions, the Antonio Nalin Center of the Baggiovara Hospital in Modena (Italy), and the Umberto I Hospital in Rome (Italy), during routine and targeted scrotal ultrasound examinations performed by expert clinicians. As a result, the dataset includes both healthy testes and images with heterogeneous pathological findings (e.g., parenchymal inhomogeneity), reflecting clinical variability. Acquisition was performed using multiple ultrasound systems: 851 images were acquired with Esaote® MyLab25 Gold and Esaote® MyLab XPro80 scanners, while 202 images were acquired with a Philips EPIQ 5 system equipped with an L12-5 linear-array transducer. Fig. 1 shows representative original images and the corresponding inputs after automatic preprocessing, which consists of Connected-Component Labeling (CCL) [6] to remove non-informative borders and acquisition overlays while preserving the anatomical field of view.

Pixel-level segmentation masks were generated using the Napari software, following a standardized protocol involving three clinicians with more than five years of experience at each institution. Each clinician annotated one third of the images from scratch and reviewed

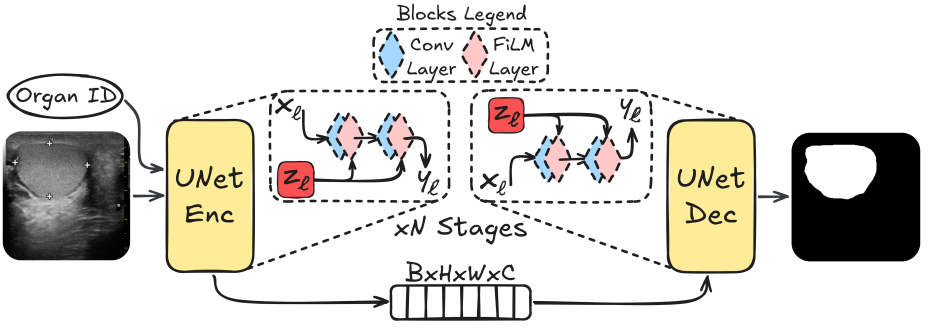

Figure 2: Overview of the proposed Cond-UNet architecture. The dashed outline marks the FiLM-modulated convolutional blocks. Conditioning vectors z_ℓ are generated from either learned organ embeddings or the proposed shared attention conditioning module.

the remaining two thirds, refining them when needed; thus, all 1,053 images were inspected by all three clinicians. Final masks were obtained by per-pixel majority voting. Pixel-wise inter-rater agreement over the full dataset is 97.2%, computed between each initial mask and the two corresponding clinician-refined masks.

The final fully anonymized dataset is released[1] with expert masks and is used in this work to support open-organ evaluation on a clinically relevant anatomical target not represented in existing public ultrasound segmentation benchmarks.

Ethical approval for the anonymized data and corresponding expert annotations was obtained from the Comitato Etico Area Vasta Emilia Nord (CE AVEN) for the TRUST-AI study (386/2025/OSS/AOUMO; approval dated December 10, 2025).

3.2 The Proposed Cond-UNet

Our segmentation model, shown in Fig. 2, builds upon a 2D U-Net architecture operating on input batched images $I \in \mathbb{R}^{B\times 3\times 512\times 512}$. We consider both four- and five-scale encoder–decoder variants. Each stage is composed of two consecutive 3×3 convolutions (with stride = 1, and zero-padding = 1) each followed by instance normalization and LeakyReLU activation. The number of feature channels in the first encoder stage is defined by the base width C_0, and channel dimensionality doubles after each downsampling step, allowing us to control model capacity through C_0. Spatial resolution is reduced via 2×2 max pooling in the encoder and restored using transposed convolutions (with kernel size = 2, stride = 2) in the decoder. Skip connections concatenate encoder features with decoder activations at corresponding resolutions. A final 1×1 convolution produces the binary segmentation map.

To enable organ-conditioned segmentation, we insert a Feature-wise Linear Modulation (FiLM) [37] layer immediately after each convolutional block of the U-Net backbone. Since each encoder and decoder stage is composed of two convolutional blocks, each stage contains two FiLM layers. This allows the conditioning signal to modulate the feature representation repeatedly at each spatial resolution, rather than only once per scale. In detail, given the feature tensor produced by the ℓ-th modulated block, $x_\ell \in \mathbb{R}^{B\times C_\ell\times H_\ell\times W_\ell}$, FiLM applies a channel-wise affine transformation:

$$y_\ell = \gamma_\ell(z_\ell)\odot x_\ell + \beta_\ell(z_\ell), \tag{1}$$

[1] https://ditto.ing.unimore.it/testiculus/

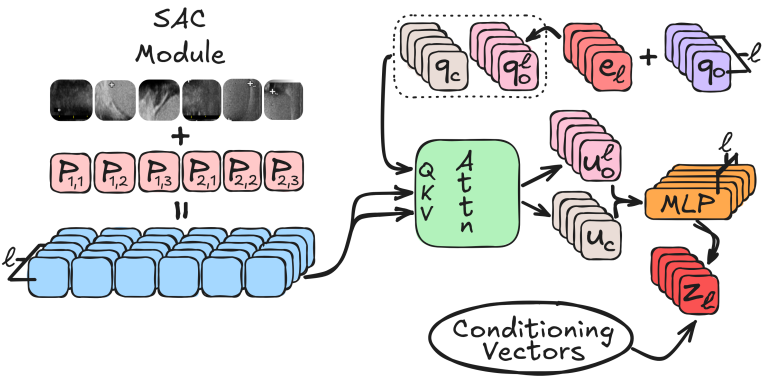

Figure 3: Overview of the proposed Cond-UNet shared attention conditioning (SAC).

where $z_\ell \in \mathbb{R}^d$ is the conditioning vector of the layer ℓ. The functions $\gamma_\ell, \beta_\ell : \mathbb{R}^d \to \mathbb{R}^{C_\ell}$ are implemented as layer-specific learnable projections that map z_ℓ from the conditioning space to the feature-channel space. Thus, they produce C_ℓ scaling and shifting coefficients, one for each feature channel, which are then broadcast over the spatial dimensions $H_\ell \times W_\ell$ and applied to x_ℓ. These projections are shared in design across all conditioning variants; therefore, the role of each conditioning mechanism is only to define how the layer-specific vector z_ℓ is computed. FiLM layers are initialized such that $\gamma_\ell \approx \mathbf{1}$ and $\beta_\ell \approx \mathbf{0}$, resulting in an approximate identity transformation at the beginning of training. This allows the network to start from the behavior of the underlying U-Net and progressively learn organ-dependent feature modulation.

We explore two alternative strategies to compute z_ℓ: (*i*) embedding-based conditioning, which depends only on the requested organ identity, and (*ii*) our proposed shared attention conditioning (SAC), which computes input-adaptive vectors from image patch embeddings.

Embedding-based conditioning (EC). The EC variant computes z_ℓ only from the requested organ identity, using a learned organ embedding shared across all FiLM layers. Given an organ label $o \in \{0, \ldots, N_{\text{org}} - 1\}$, a learnable organ embedding $e_o \in \mathbb{R}^{64}$ is retrieved from an organ embedding table initialized from $\mathcal{N}(0, 0.02)$. This vector provides a compact representation of the target anatomy and is shared across all FiLM layers. Since different network depths require different modulation parameters, each FiLM layer ℓ has its own MLP f_ℓ, which maps the organ embedding to a layer-specific conditioning vector:

$$z_\ell = f_\ell(e_o). \tag{2}$$

The resulting vector z_ℓ is then used to generate the FiLM scaling and shifting coefficients for layer ℓ according to Eq. (1). In this way, the same organ identity can induce distinct modulations at different network depths. However, because z_ℓ depends only on the organ embedding e_o, this variant applies a fixed organ-specific modulation pattern for all images of the same target organ, independently of the actual image appearance.

Shared attention conditioning (SAC). The SAC variant computes z_ℓ from both the requested organ identity and the input, as depicted in Fig. 3. The image I is first divided into P patches, projected into embeddings used as keys and values $K, V \in \mathbb{R}^{B \times P \times D}$. These patch embeddings provide the image content used by the conditioning module.

For each target organ, SAC learns an organ query $q_o \in \mathbb{R}^D$. To make this query specific to each FiLM layer, we add a learnable layer embedding $e_\ell \in \mathbb{R}^D$ and define $q_o^\ell = q_o + e_\ell$. In parallel, SAC also learns a compression query $q_c \in \mathbb{R}^D$, initialized in the same way as the organ queries, to extract an organ-agnostic global image summary. Both queries attend to the image patch embeddings through a shared multi-head attention module $\mathcal{A}(\cdot)$:

$$u_o^\ell = \mathcal{A}(q_o^\ell, K, V), \qquad u_c = \mathcal{A}(q_c, K, V). \tag{3}$$

The vector u_o^ℓ can be interpreted as an image-adaptive summary of the regions and appearance patterns relevant to organ o at layer ℓ, while u_c captures global image information independent of the target organ. The organ-specific and global summaries are then averaged and passed through a layer-specific MLP, yielding the conditioning vector:

$$z_\ell = f_\ell\left(\frac{u_o^\ell + u_c}{2}\right). \tag{4}$$

Unlike embedding-based conditioning, SAC does not directly map an organ identity to FiLM modulation parameters. Instead, the organ embedding acts as a layer-specific attention query over image patches, extracting organ-relevant image information that is then used to generate the FiLM conditioning vector z_ℓ. Thus, z_ℓ depends jointly on the target organ, the image content, and the FiLM layer, enabling image-adaptive rather than fixed organ-specific modulation. Since SAC operates only on input patch embeddings and not on intermediate U-Net activations, the conditioning vectors for all FiLM layers can be computed in parallel.

4 Experimental Settings

A central part of our study is a comprehensive benchmark of ultrasound segmentation methods across different organs and different centers, as well as open-organ settings. We trained and finetuned a heterogeneous set of segmentation models on the UUSIC dataset to evaluate cross-organ segmentation performance and generalization across acquisition centers. In addition, we use our multicenter testicular ultrasound dataset to assess transfer to an unseen anatomical target acquired with different ultrasound machines.

4.1 Datasets and Evaluation Protocol

We use the UUSIC [24] dataset as the main source for training and cross-organ multicenter evaluation. UUSIC aggregates multiple ultrasound datasets covering both classification and binary segmentation tasks. Since our work focuses on pixel-wise segmentation, we retain only the images with segmentation ground-truth masks, resulting in $6,613$ images.

Among these, $5,535$ images belong to publicly available datasets, i.e., UUSIC-Public, which we denote as *UUSIC-P*. This subset includes breast nodules from BUSI, BUSIS [56], and BUS-BRA [10], thyroid nodules from DDTI [36], kidney images from KidneyUS [40], fetal head images from FetalHead [42], and cardiac chamber images from CAMUS [22]. The remaining $1,078$ images were internally acquired by the UUSIC challenge organizers, i.e., UUSIC-Internal, and are denoted as *UUSIC-I*. This private subset mirrors the anatomical categories of *UUSIC-P*, providing an independent counterpart for each organ class and enabling evaluation of cross-center generalization within the same anatomical targets. Notably, the private data have not been used for training existing foundation models.

In our protocol, all models are trained and validated only on *UUSIC-P*. We then evaluate them on *UUSIC-I* to measure robustness to domain shifts across acquisition centers. Finally, we use the TesticulUS-Real dataset, containing 1,053 real images acquired with two different ultrasound systems, to assess open-organ transfer. Since no public testicular US segmentation dataset previously existed, testicular US represents an unseen anatomical target for both trained and pretrained models. We therefore evaluate models directly on this organ, without testicular training examples, to measure cross-organ generalization.

4.2 Model Implementation

To ensure a rigorous and fair comparison, we selected baselines spanning fully-supervised task-specific models (nnU-Net [17], Swin-UNETR [12], and U-Mamba [30]), a semi-supervised ultrasound segmentation model (SPSS [8]), a model designed for ultrasound organ generalization (MI-SegNet [5]), and SAM-based foundation models, pretrained on several types of medical imaging (MedSAM [29]) or specifically pretrained on ultrasound (UltraSAM [32] and AutoSAMUS [23]). Each method was trained to match our experimental setting in terms of training data, input resolution, and evaluation split. Our proposed model and all the competitors were trained on NVIDIA L40S GPUs with 45 GB of VRAM, using full-precision arithmetic. For each architectural variant, experiments were repeated across three seeds (42, 78, 123), and the mean and standard deviation are reported over these runs. Each run was trained until convergence, and the checkpoint that achieved the lowest validation loss was selected for the final evaluation.

Our model. Our Cond-UNet was optimized using AdamW with an initial learning rate of 10^{-3} and cosine annealing scheduling. Training was performed for 10,000 optimization steps with a batch size of 4 and gradient accumulation over 4 mini-batches, yielding an effective batch size of 16. The segmentation loss consists of a standard combination of Dice loss and binary cross-entropy.

Task-specific models. For the fully supervised task-specific baselines, we selected one representative method for each main architectural family: nnU-Net [17] as a convolutional model, Swin-UNETR [12] as a Transformer-based model, and U-Mamba [30] as a Mamba-based model. For nnU-Net, we followed its self-configuring pipeline, where preprocessing, architecture, optimization, and training settings are automatically adapted to the dataset characteristics. The only imposed constraint was the available GPU memory budget, which was matched to the resources used in our experiments. For Swin-UNETR and U-Mamba, we integrated the corresponding architectures within the nnU-Net framework and tuned the main training and memory-related hyperparameters to make full use of the available GPU memory while keeping the same data splits and preprocessing pipeline. We also included SPSS [8] as a semi-supervised ultrasound segmentation baseline, reproducing its architecture and training strategy according to the original paper. Finally, we evaluated MI-SegNet [5], which is specifically designed for unseen-domain generalization in ultrasound segmentation, with only minimal changes required to match our segmentation setting and dataset organization.

Foundation models. We evaluated MedSAM [29], UltraSAM [32], and AutoSAMUS [23] as representative SAM-derived foundation models for medical and ultrasound image segmentation. Since these methods start from large pretrained checkpoints, we adapted them using two finetuning regimes: a parameter-efficient LoRA-based [14] strategy, following common SAM adaptation protocols for medical segmentation [52, 57], and full end-to-end finetuning of the trainable weights. The LoRA setting was intended to preserve the pre-

trained representation while adapting the models with a limited number of trainable parameters. To avoid annotation-derived prompts at inference time, all SAM-based models were evaluated with fully automatic prompting. For MedSAM and UltraSAM, we learned organ-specific prompt embeddings compatible with their prompt encoders. For AutoSAMUS, we initialized the model with the authors' released checkpoint and used its learned automatic prompting mechanism while finetuning on UUSIC-P. In all cases, inference was performed without expert interaction or ground-truth-derived prompts.

5 Results and Discussion

5.1 Cross-Organ Generalization

Tab. 1 reports the cross-organ generalization performance on UUSIC-I, with all models trained or finetuned on UUSIC-P. We include task-specific baselines, SAM-based foundation models, and the two variants of our conditional network, i.e., Cond-UNet with embedding-based conditioning (EC) and Cond-UNet with shared attention conditioning (SAC). Since the anatomical categories are preserved between train and test splits, this setting mainly evaluates robustness to acquisition and institutional shifts across organs.

A clear separation emerges between standard task-specific models and pretrained foundation models. In most cases, large-scale pretraining (see Tab. 3 for more details) provides a measurable advantage under domain shift, with MedSAM, UltraSAM, and AutoSAMUS generally outperforming nnU-Net, Swin-UNETR, U-Mamba, and MI-SegNet. The main exception is SPSS, which is the strongest task-specific competitor, suggesting that incorporating shape priors through a semi-supervised objective can partially compensate for the absence of large-scale pretraining. Nevertheless, both variants of our method outperform all competitors on average across the five UUSIC-I organs, with Cond-UNet (EC) reaching 77.2 ± 0.7 Dice and Cond-UNet (SAC) achieving the best overall mean of 78.3 ± 0.1. SAC performs best on breast and cardiac and second best on kidney, while EC achieves the best kidney performance, indicating robust performance across organs. Both of them also remain competitive on fetal and thyroid images, leading to a more balanced performance profile than the foundation-model baselines. This suggests that organ-aware conditioning can improve robustness without relying on large-scale pretraining or prompt-based inference, possibly because the conditioning signal is largely domain-agnostic. Although image appearance changes across centers (because of scanner, operator, and acquisition differences), the requested anatomical target remains fixed, encouraging the shared backbone to focus on organ-consistent structures rather than center-specific intensity or texture cues.

Thyroid is the only class where UltraSAM and AutoSAMUS outperform both Cond-UNet variants, likely reflecting the benefit of large-scale pretraining for anatomical regions with limited supervised examples in the training split. However, this advantage is not consistent across organs, as the same models show weaker performance on other organs, most notably cardiac segmentation.

Overall, Tab. 1 shows that foundation models can help on specific organs, especially in low-data regimes, but are not consistently superior across all anatomical targets. Task-specific models remain competitive when combined with generalization-oriented objectives, as shown by SPSS. In this setting, Cond-UNet achieves the most consistent cross-organ performance through a lightweight fully automatic backbone with organ-aware modulation.

Table 1: Comparison of selected models on cross-domain generalization. Dice scores (%) are reported as mean ± std. Results for the Testicle column are obtained in a strict open-organ setting on the newly introduced testicular ultrasound dataset (TesticulUS-Real), without any model adaptation. The best results are in **bold** while the second best are underlined. Dagger denotes LoRA finetuned models.

	Model	Breast	Cardiac	Fetal	Kidney	Thyroid	Mean	Testicle
Task-specific	nnU-Net	77.9 ± 0.2	74.3 ± 4.1	90.7 ± 0.2	88.0 ± 0.3	32.2 ± 2.8	72.6 ± 0.8	64.7 ± 6.1
	Swin-UNETR	72.9 ± 3.4	71.6 ± 1.0	90.6 ± 0.9	87.4 ± 0.8	32.5 ± 3.0	71.0 ± 1.7	58.1 ± 11.4
	U-Mamba	74.1 ± 3.6	67.1 ± 2.5	89.4 ± 1.3	88.6 ± 0.1	29.7 ± 2.9	69.8 ± 0.6	74.8 ± 6.9
	SPSS	78.6 ± 0.9	70.7 ± 3.3	**91.9 ± 0.1**	86.5 ± 0.9	49.3 ± 2.2	75.4 ± 0.3	60.2 ± 3.4
	MI-SegNet	60.4 ± 1.0	71.1 ± 1.5	86.8 ± 1.0	83.1 ± 1.0	32.4 ± 1.9	66.8 ± 0.5	62.3 ± 2.6
Foundation	MedSAM	74.3 ± 0.3	72.3 ± 0.7	89.7 ± 0.8	88.2 ± 0.4	51.6 ± 0.9	75.3 ± 0.2	72.1 ± 6.2
	MedSAM†	77.5 ± 0.8	69.9 ± 2.1	91.6 ± 0.2	88.1 ± 0.7	51.5 ± 0.7	75.7 ± 0.5	76.3 ± 5.6
	UltraSAM	79.7 ± 0.7	58.9 ± 0.9	86.2 ± 0.9	85.5 ± 0.4	58.7 ± 1.0	73.8 ± 0.4	**79.8 ± 1.2**
	UltraSAM†	77.6 ± 1.1	62.3 ± 5.0	86.8 ± 0.1	85.4 ± 1.9	58.1 ± 1.2	74.1 ± 1.4	78.4 ± 2.0
	AutoSAMUS	73.6 ± 1.2	57.9 ± 0.7	86.7 ± 0.5	84.1 ± 0.8	**59.2 ± 1.4**	72.3 ± 0.4	77.1 ± 3.0
	AutoSAMUS†	75.4 ± 0.9	56.9 ± 1.3	88.0 ± 1.5	84.3 ± 1.3	55.2 ± 2.0	72.0 ± 0.9	75.8 ± 4.2
Ours	**Cond-UNet (EC)**	79.4 ± 1.1	76.9 ± 0.9	89.6 ± 0.4	**89.4 ± 0.4**	50.5 ± 1.9	77.2 ± 0.7	71.8 ± 5.9
	Cond-UNet (SAC)	**80.6 ± 0.4**	**79.7 ± 1.2**	89.9 ± 0.2	89.2 ± 0.3	51.9 ± 1.3	**78.3 ± 0.1**	77.3 ± 6.7

5.2 Open-Organ Evaluation

The last column of Tab. 1 reports performance on our TesticulUS-Real dataset, which we use to define an open-organ evaluation setting. This protocol measures the ability of each model to generalize to an anatomical region that is absent from the training set. The results reflect a clear trend: models with larger capacity and extensive pretraining (details in Tab. 3) generally achieve stronger generalization to this unseen organ than task-specific architectures trained only on UUSIC-P. This is expected, since large pretrained models have been exposed to substantially richer visual variability before adaptation, whereas methods such as nnU-Net, SPSS, and MI-SegNet must infer the unseen anatomy only from the limited supervised multi-organ training set.

For our conditioned model, open-organ inference requires querying the network with an organ identity that was never observed during training. We therefore sample a new embedding from the same Gaussian initialization distribution used for the original organ embeddings, $\mathcal{N}(0, 0.02)$. This embedding is not trained and does not encode any prior information about the testicle; it only provides a valid input to the conditioning pathway. Thus, the prediction mainly reflects the model's learned ability to segment foreground structures in ultrasound under an unseen anatomical request, rather than memorization of an organ-specific code. For each inference run, a single sampled embedding is fixed and used for all testicular test images. To assess robustness to the sampled embedding, we repeated zero-shot inference with five independently sampled tokens, obtaining Dice scores of 71.8 ± 0.1 for EC and 77.4 ± 0.1 for SAC, indicating low sensitivity to the particular conditioning vector. Under this strict setting, Cond-UNet (SAC) is the second best method after UltraSAM (despite using far fewer parameters), achieving much better performance than task-specific models and all the other foundation models.

Table 2: Ablation results (Dice %) evaluating model scale (C_0, Depth) and conditioning mechanism (No = unconditioned baseline). Values are mean ± std over three seeds. Parameters are reported in millions. Best results are reported in **bold**, second best are underlined.

C_0	D	Cond.	Breast	Cardiac	Fetal	Kidney	Thyroid	Mean	# Params
8	4	No	69.7 ± 0.9	64.8 ± 3.2	87.1 ± 2.0	87.8 ± 0.3	47.7 ± 0.6	71.4 ± 1.1	1.8
		EC	76.4 ± 1.1	73.7 ± 0.7	88.7 ± 0.6	87.9 ± 0.6	47.2 ± 1.3	74.8 ± 0.4	2.6
		SAC	77.0 ± 0.5	76.6 ± 2.4	89.7 ± 0.1	88.6 ± 0.2	46.8 ± 1.5	75.7 ± 0.4	18.5
16	4	No	69.5 ± 0.9	61.7 ± 1.3	87.6 ± 0.4	88.3 ± 0.9	47.5 ± 0.7	70.9 ± 0.5	7.0
		EC	77.4 ± 0.6	75.7 ± 2.1	89.7 ± 0.2	88.9 ± 0.5	47.1 ± 1.3	75.8 ± 0.7	9.8
		SAC	78.3 ± 0.2	77.3 ± 0.5	**90.7 ± 0.1**	89.4 ± 0.2	47.2 ± 1.8	76.6 ± 0.4	24.2
8	5	No	71.7 ± 0.4	72.7 ± 3.6	89.5 ± 1.0	88.9 ± 0.1	51.6 ± 0.3	74.9 ± 1.0	7.0
		EC	78.9 ± 0.3	76.9 ± 0.7	89.8 ± 0.2	89.4 ± 0.3	50.6 ± 1.4	77.1 ± 0.3	9.8
		SAC	**81.3 ± 0.4**	77.5 ± 0.1	90.4 ± 0.6	**89.6 ± 0.2**	50.6 ± 1.0	77.9 ± 0.2	26.5
16	5	No	69.6 ± 0.1	62.9 ± 3.9	88.5 ± 0.2	88.6 ± 0.2	51.8 ± 1.9	72.3 ± 0.7	27.8
		EC	79.4 ± 1.1	76.9 ± 0.9	89.6 ± 0.4	89.4 ± 0.4	50.5 ± 1.9	77.2 ± 0.7	38.3
		SAC	80.6 ± 0.4	**79.7 ± 1.2**	89.9 ± 0.2	89.2 ± 0.3	**51.9 ± 1.3**	**78.3 ± 0.1**	48.1

The strong performance of UltraSAM on the testicular dataset should be interpreted in light of its substantially larger ultrasound exposure during pretraining as shown in Tab. 3. This difference in ultrasound-specific data exposure suggests that UltraSAM's open-organ advantage is likely driven in part by ultrasound-scale pretraining, rather than by the evaluation protocol alone.

Table 3: Medical images (US where specified) seen during pretraining.

Model	# Images
MedSAM	1.5M (95K US)
AutoSAMUS	~30,000 US
UltraSAM	282,321 US
Others	-

5.3 Ablation on Model Scale and Conditioning

Tab. 2 evaluates the effect of the conditioning strategy and model capacity. Across all configurations, adding FiLM-based conditioning improves over the unconditioned baseline, confirming the benefit of organ-aware modulation for robust segmentation. Embedding-based conditioning (EC) already provides substantial gains, while replacing the fixed organ embedding with attention-based conditioning (SAC) consistently achieves the best performance within each capacity group. The table also shows that conditioning is more important than scale alone. At the same time, increasing depth and width further improves performance when combined with conditioning, with SAC giving the most stable gains across organs.

Fig. 4 provides empirical insight into how the two conditioning mechanisms modulate the network, by visualizing the mean FiLM scaling parameters γ across the 22 FiLM layers, grouped by organ. With EC, modulation is concentrated in a limited set of intermediate layers and follows relatively similar patterns across organs, suggesting that the organ embedding mainly acts as a coarse global bias, with limited control over high-resolution layers. In contrast, SAC produces more distributed and organ-dependent modulation across both encoder and decoder layers, indicating that image-adaptive attention generates conditioning signals that vary with both anatomical target and network depth.

To assess the functional relevance of these modulation patterns, we perform a layer-wise FiLM ablation by replacing the modulation of one layer at a time with the identity

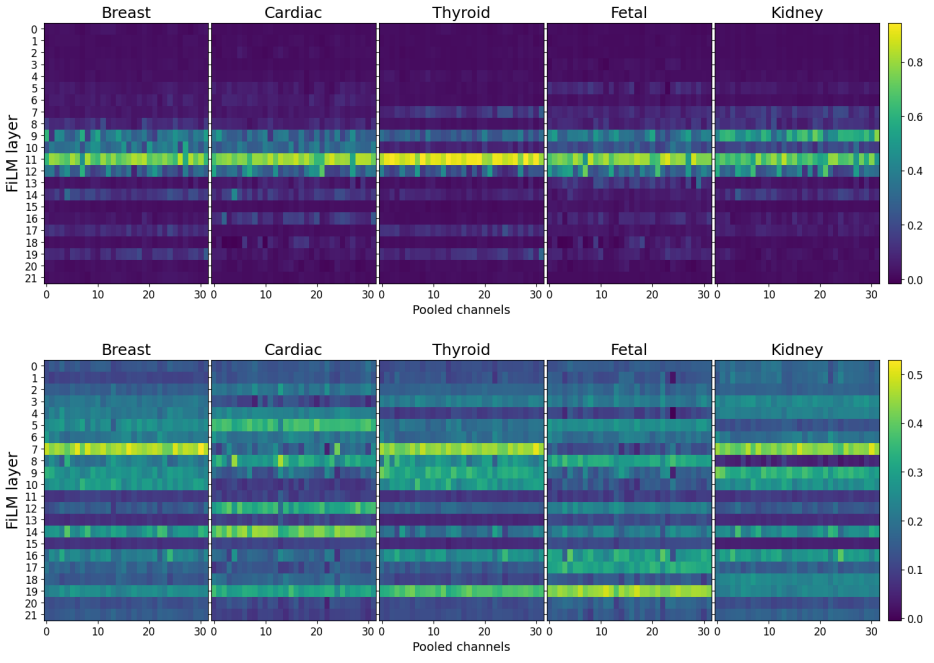

Figure 4: Visualization of the mean FiLM scaling parameters γ across the 22 FiLM layers, grouped by organ. The first row corresponds to Cond-UNet with embedding-based conditioning, while the second row corresponds to SAC conditioning.

transformation, i.e., γ=1 and β=0. The EC model shows only moderate Dice degradation for most layers, consistent with a relatively redundant conditioning signal concentrated in a small subset of the network. SAC, instead, exhibits larger drops, particularly in the deeper decoder and high-resolution reconstruction stages. This suggests that SAC not only produces more diverse modulation but also assigns functional importance to selected layers, enabling the network to exploit organ- and image-dependent conditioning at multiple resolutions.

These observations align with Tab. 2, where SAC achieves the highest mean Dice within each capacity group, with the largest model performing best overall.

5.4 Efficiency Analysis

We further analyze model efficiency in terms of parameter count, computational cost, and inference time, important in this scenario for practical deployability. Fig. 5 reports the Pareto analysis between mean Dice and GFLOPs, while Fig. 6 jointly compares mean Dice, parameter count, and computational cost. The results show that increasing model size does not necessarily lead to better segmentation performance. Several task-specific and foundation-model baselines require substantially more parameters or GFLOPs, yet do not consistently outperform compact alternatives. In contrast, multiple Cond-UNet variants lie on or close to the Pareto frontier, indicating a more favorable accuracy-complexity trade-off. Tab. 4, which reports our and the most efficient models in terms of inference time of each subcategory (i.e., U-Mamba, and MedSAM), shows that both Cond-UNet variants are faster and use less VRAM, supporting parallel inference even on lower-memory GPUs.

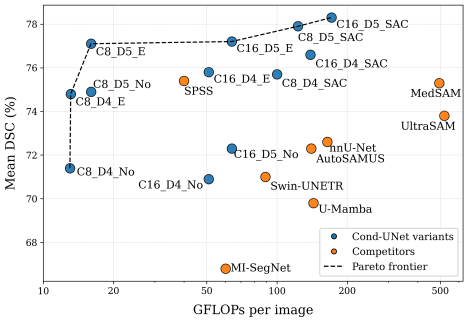

Figure 5: Pareto efficiency analysis comparing Cond-UNet variants with competing methods. The dashed line indicates the Pareto frontier. Model names follow Table 2, but with E denoting EC variant.

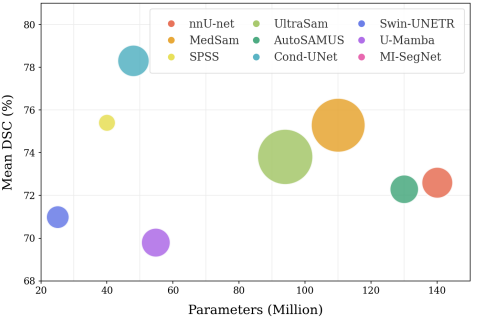

Figure 6: Bubble plot illustrating the relationship between segmentation performance (mean Dice, %) and model size. The bubble area represents the computational cost measured in GFLOPs per image.

This trend is also consistent with the ablation results in Tab. 2. Compact conditioned models already achieve strong Dice scores. Notably, the embedding-conditioned configuration with only 2.6M parameters reaches a mean Dice of 74.8, outperforming larger unconditioned variants. This suggests that the benefit of our approach comes not simply from model capacity, but from a more effective use of parameters through organ-aware modulation. SAC further improves this behavior by introducing input-adaptive conditioning, achieving the best mean Dice while remaining considerably smaller than foundation models.

Overall, these results show that architectural efficiency and conditioning strategy are as important as raw model scale for cross-organ segmentation. Rather than relying only on large pretrained backbones, Cond-UNet provides a computationally efficient alternative with strong segmentation performance on ultrasound images.

Table 4: Single-image inference time on a 16GB TESLA V100, peak reserved GPU memory, number of parameters and GFLOPs.

Model	Time (ms)	Peak VRAM	# Params	GFLOPs
Our (EC)	19.04	392 MB	38.3 M	64.0
Our (SAC)	21.13	511 MB	48.1 M	171.0
U-Mamba	60.54	667 MB	54.8 M	143.1
MedSAM	120.96	3206 MB	110.0 M	495.0

6 Conclusions

In this work, we introduced TesticulUS-Real, a new multicenter testicular ultrasound segmentation dataset, providing the first public real-image resource with expert masks for this anatomical target. Using this dataset together with UUSIC, we defined a benchmark for multicenter cross-organ generalization and open-organ ultrasound segmentation. We also proposed Cond-UNet, a lightweight organ-conditioned architecture based on FiLM modulation and shared attention conditioning. Experiments show that Cond-UNet achieves the best average cross-organ performance with a better accuracy-complexity trade-off compared with larger foundation models. In the open-organ setting, our testicular dataset highlights the role of large-scale ultrasound pretraining, while showing that our compact model can generalize competitively. Overall, our results support lightweight conditional architectures as a promising direction for robust and deployable multi-organ ultrasound segmentation.

Acknowledgements. This project is funded by the University of Modena and Reggio Emilia and Fondazione di Modena through FAR-2025 (E93C26000100007) and FARD-2024, and by the Italian Ministry of Research's NRRP complementary actions "Fit4MedRob – Fit for Medical Robotics" (PNC0000007).

References

[1] Mudassar Ali, Tong Wu, Haoji Hu, Qiong Luo, Dong Xu, Weizeng Zheng, Neng Jin, Chen Yang, and Jincao Yao. A review of the Segment Anything Model (SAM) for medical image analysis: Accomplishments and perspectives. *Computerized Medical Imaging and Graphics*, 119, 2025.

[2] Md Zahangir Alom, Chris Yakopcic, Mahmudul Hasan, Tarek M Taha, and Vijayan K Asari. Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging*, 6(1), 2019.

[3] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation Models Defining a New Era in Vision: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4), 2025.

[4] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical Image Segmentation Review: The Success of U-Net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 2024.

[5] Yuan Bi, Zhongliang Jiang, Ricarda Clarenbach, Reza Ghotbi, Angelos Karlas, and Nassir Navab. MI-SegNet: Mutual information-based US segmentation for unseen domain generalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

[6] Federico Bolelli, Stefano Allegretti, Luca Lumetti, and Costantino Grana. A State-of-the-Art Review With Code About Connected Components Labeling on GPUs. *IEEE Transactions on Parallel and Distributed Systems*, 37(4), 2026.

[7] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97, 2024.

[8] Yaxiong Chen, Yujie Wang, Zixuan Zheng, Jingliang Hu, Yilei Shi, Shengwu Xiong, Xiao Xiang Zhu, and Lichao Mou. Striving for Simplicity: Simple Yet Effective Prior-Aware Pseudo-labeling for Semi-supervised Ultrasound Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[9] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. SAM-Med2D. *arXiv preprint arXiv:2308.16184*, 2023.

[10] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4), 2024.

[11] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[12] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In *International MICCAI Brainlesion Workshop*, 2021.

[13] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

[15] Qinghua Huang, Fan Zhang, and Xuelong Li. Machine Learning in Ultrasound Computer-Aided Diagnostic Systems: A Survey. *BioMed Research International*, 2018 (1), 2018.

[16] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 2021.

[17] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[18] Asifullah Khan, Zunaira Rauf, Abdul Rehman Khan, Saima Rathore, Saddam Hussain Khan, Najamus Saher Shah, Umair Farooq, Hifsa Asif, Aqsa Asif, Umme Zahoora, Rafi Ullah Khalil, Suleman Qamar, Umme Hani Tayyab, Faiza Babar Khan, Abdul Majid, and Jeonghwan Gwak. A Recent Survey of Vision Transformers for Medical Image Segmentation. *IEEE Access*, 13, 2025.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[20] Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Prateek Prasanna, Parminder Bhatia, and Taha Kass-Hout. Enhancing SAM with Efficient Prompting and Preference Optimization for Semi-supervised Medical Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[21] Yogiek Indra Kurniawan, Muhammad Febrian Rachmadi, Alif Wicaksana Ramadhan, and Wisnu Jatmiko. Mamba-Based Deep Learning Methods in Medical Image Analysis: A Systematic Literature Review. *IEEE Access*, 13, 2025.

[22] Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Lovstakken, and Olivier Bernard. Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. *IEEE Transactions on Medical Imaging*, 38(9), 2019.

[23] Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting SAM: Towards End-to-End Ultrasound Image Segmentation Via Auto Prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[24] Zehui Lin, Luyi Han, Tianyu Zhang, Xing Wang, Ying Zhou, Yanming Zhang, Dong Xu, Lingyun Bao, Ritse Mann, Shuo Li, Shandong Wu, Dong Ni, and Tao Tan. Universal Ultrasound Image Challenge: Multi-Organ Classification and Segmentation (UUSIC25), 2025. URL https://doi.org/10.5281/zenodo.15094669.

[25] Xiao Liu, Chenxu Zhang, Fuxiang Huang, Shuyin Xia, Guoyin Wang, and Lei Zhang. Vision Mamba: A Comprehensive Survey and Taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, 37(2), 2026.

[26] Francesco Lotti, Michal Studniarek, Cristina Balasa, Jane Belfield, Pieter De Visschere, Simon Freeman, Oliwia Kozak, Karolina Markiet, Subramaniyan Ramanathan, Jonathan Richenberg, Mustafa Secil, Katarzyna Skrobisz, Athina C. Tsili, Michele Bertolotto, and Laurence Rocher. The role of the radiologist in the evaluation of male infertility: recommendations of the european society of urogenital radiology-scrotal and penile imaging working group (esur-spiwg) for scrotal imaging. *European Radiology*, 35(2), 2025.

[27] Luca Lumetti, Vittorio Pipoli, Kevin Marchesini, Elisa Ficarra, Costantino Grana, and Federico Bolelli. Taming Mambas for 3D Medical Image Segmentation. *IEEE Access*, 13, 2025.

[28] Luca Lumetti, Vittorio Pipoli, Kevin Marchesini, Elisa Ficarra, Costantino Grana, and Federico Bolelli. Accurate 3D Medical Image Segmentation with Mambas. In *IEEE International Symposium on Biomedical Imaging*, 2025.

[29] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), 2024.

[30] Jun Ma, Feifei Li, and Bo Wang. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[31] Zengchen Ma, Yanwen Sun, Shixun Zhai, Guocheng Lei, and Kaige Zhang. A Review of Segment Anything Model and Its Applications. In *2024 IEEE International Conference on Unmanned Systems*, 2024.

[32] 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*, 21(1), 2026.

[33] Nicola Morelli, Kevin Marchesini, Luca Lumetti, Daniele Santi, Costantino Grana, and Federico Bolelli. Enhancing Testicular Ultrasound Image Classification Through Synthetic Data and Pretraining Strategies. In *International Conference on Image Analysis and Processing*, 2025.

[34] Muhammad Nouman, Ghada Khoriba, and Essam A Rashed. Rethinking MedSAM: Performance Discrepancies in Clinical Applications. In *IEEE International Conference on Future Machine Learning and Data Science*, 2024.

[35] World Health Organization. *Infertility prevalence estimates, 1990–2021*. World Health Organization, 2023.

[36] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *International Symposium on Medical Information Processing and Analysis*, 2015.

[37] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI Conference on Artificial Intelligence*, 2018.

[38] Vittorio Pipoli, Alessia Saporita, Kevin Marchesini, Costantino Grana, Elisa Ficarra, and Federico Bolelli. IM-Fuse: A Mamba-Based Fusion Block for Brain Tumor Segmentation with Incomplete Modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2025.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[40] Rohit Singla, Cailin Ringstrom, Grace Hu, Victoria Lessoway, Janice Reid, Christopher Nguan, and Robert Rohling. The Open Kidney Ultrasound Data Set. In *International Workshop on Advances in Simplifying Medical Ultrasound*, 2023.

[41] Giorgia Spaggiari, Antonio RM Granata, and Daniele Santi. Testicular ultrasound inhomogeneity is an informative parameter for fertility evaluation. *Asian Journal of Andrology*, 22(3), 2020.

[42] Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS One*, 13(8), 2018.

[43] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, and Junjun He. SAM-Med3D: A Vision Foundation Model for General-Purpose Segmentation on Volumetric Medical Images. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10), 2025.

[44] Jinhong Wang, Jintai Chen, Danny Chen, and Jian Wu. LKM-UNet: Large Kernel Vision Mamba UNet for Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[45] Yu Wang, Xinke Ge, He Ma, Shouliang Qi, Guanjing Zhang, and Yudong Yao. Deep Learning in Medical Ultrasound Image Analysis: A Review. *IEEE Access*, 9, 2021.

[46] Zhikai Wei, Chao Wu, Hanyu Du, Rui Yu, Bo Du, and Yongchao Xu. Noise-Robust Tuning of SAM for Domain Generalized Ultrasound Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2025.

[47] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis*, 102, 2025.

[48] Qingling Xia, Hong Zheng, Haonan Zou, Dinghao Luo, Hongan Tang, Lingxiao Li, and Bin Jiang. A comprehensive review of deep learning for medical image segmentation. *Neurocomputing*, 613, 2025.

[49] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[50] Danping Yin, Qingqing Zheng, Long Chen, Ying Hu, and Qiong Wang. APG-SAM: Automatic prompt generation for SAM-based breast lesion segmentation with boundary-aware optimization. *Expert Systems with Applications*, 276, 2025.

[51] Jiawei Zhang, Yanchun Zhang, Yuzhen Jin, Jilan Xu, and Xiaowei Xu. Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation. *Health Information Science and Systems*, 11(1), 2023.

[52] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[53] Ruixuan Zhang, Wenhuan Lu, Cuntai Guan, Jie Gao, Xi Wei, and Xuewei Li. SHAN: Shape Guided Network for Thyroid Nodule Ultrasound Cross-Domain Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[54] Weijie Zhang, Lingfeng Xie, Kun Zeng, Xiaonan Luo, and Yongyi Gong. DPGS-Net: Dual Prior-Guided Cross-Domain Adaptive Framework for Ultrasound Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2025.

[55] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171, 2024.

[56] Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. BUSIS: A Benchmark for Breast Ultrasound Image Segmentation. *Healthcare*, 10(4), 2022.

[57] Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution Meets LoRA: Parameter Efficient Finetuning for Segment Anything Model. In *International Conference on Learning Representations*, 2024.

[58] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer. *IEEE Transactions on Image Processing*, 32, 2023.