

TOPOLOGY-INFORMED PROMPT-CONDITIONED UNIVERSAL SEGMENTATION OF UTERINE STRUCTURES FROM ULTRASOUND AND MRI

Yongheng Sun¹, Yuexi Gu¹, Jingwen Sun¹, Maureen Kohi^{1,*}, Mingxia Liu^{1,*}

¹Department of Radiology and BRIC, UNC at Chapel Hill, Chapel Hill, NC 27599, USA

ABSTRACT

Multi-structure segmentation of the uterus is important for computer-assisted screening, diagnosis, and treatment planning of uterine diseases, where ultrasound and MRI provide complementary clinical information. However, developing a unified model across these modalities is challenging due to their substantially different image appearances, anatomical contexts, spatial resolutions, and label spaces. Moreover, existing datasets often define different segmentation targets, making joint learning challenging and potentially leading to negative transfer across heterogeneous tasks. To this end, we propose a Topology-informed Prompt-conditioned Universal Segmentation (TPUS) framework for segmenting multiple uterine structures across ultrasound and MRI. TPUS introduces a graph-based multi-dataset backbone comprising modality-specific stems and a modality-shared graph-based encoder-decoder to support modality-sensitive input adaptation, structural feature reasoning, and joint representation learning across heterogeneous uterine segmentation tasks. In addition, TPUS uses task-aware class prompts to condition the segmentation process for different datasets and label spaces, a dynamic convolutional adaptation module to generate task-specific output responses, and a topology-informed loss to encourage anatomically consistent predictions. Experiments on a uterine ultrasound dataset and a T2-weighted uterine myoma MRI dataset demonstrate that TPUS achieves Dice scores of 0.898 and 0.693 on the two held-out test sets, respectively, outperforming several generic and universal segmentation baselines. Source code can be accessed [online](#).

Index Terms— Uterine Structure Segmentation, Ultrasound, Magnetic Resonance Imaging

1. INTRODUCTION

Multi-structure segmentation of the uterus is an important task in medical image analysis, supporting computer-assisted screening, diagnosis, disease characterization, and treatment planning for a wide range of uterine diseases [1, 2]. Accurate delineation of the uterus and its internal structures, such as the uterine cavity and myometrium, as well as focal lesions including uterine fibroids and Nabothian cysts, can provide quantitative anatomical and pathological information for clinical assessment [2]. Ultrasound and MRI are two commonly used imaging modalities for evaluating uterine anatomy and pathology [1, 2], offering complementary information regarding accessibility, spatial detail, soft-tissue contrast, and tissue characterization. Developing robust methods to segment multiple uterine structures and lesions across ultrasound and MRI could facilitate comprehensive automated analysis of uterine diseases.

However, achieving unified segmentation across ultrasound and MRI remains challenging due to substantial differences between the

two modalities in image appearance, anatomical context, spatial resolution, and tissue contrast [1, 2]. Beyond these modality-specific differences, publicly available datasets are often designed for different clinical objectives and therefore provide heterogeneous segmentation targets and label definitions. For example, one dataset may focus on delineating the uterus and its internal structures, whereas another may primarily target uterine lesions such as fibroids. Consequently, directly combining such datasets for joint model training can introduce substantial task heterogeneity and lead to negative transfer, where knowledge beneficial to one segmentation task may interfere with another. Most existing methods are designed for a single modality, dataset, or fixed label space, and thus are not well suited for jointly learning heterogeneous uterine segmentation tasks across ultrasound and MRI.

To address these challenges, we propose a Topology-informed Prompt-conditioned Universal Segmentation (TPUS) framework for unified multi-structure segmentation across uterine ultrasound and MRI. As shown in Fig. 1, TPUS is designed to jointly learn representations across heterogeneous datasets while adapting the segmentation process to dataset-specific label spaces. Specifically, we introduce task-aware class prompts to explicitly condition the model on different segmentation targets, enabling flexible prediction across heterogeneous label definitions. A dynamic convolutional adaptation module further generates task-specific output responses, allowing the representation to be effectively adapted to different modalities and segmentation tasks. To preserve anatomical consistency, we incorporate a topology-informed loss that encourages structurally plausible predictions. In addition, we develop a graph-based multi-dataset backbone comprising modality-specific stems and a modality-shared graph-based encoder-decoder, enabling modality-sensitive input adaptation and joint representation learning across heterogeneous uterine segmentation tasks. These components enable TPUS to jointly learn from ultrasound and MRI while mitigating negative transfer and improving multi-structure uterine segmentation. Experiments on uterine ultrasound and T2-weighted MRI datasets demonstrate that the proposed TPUS consistently outperforms several generic and universal segmentation methods.

2. METHODOLOGY

The proposed TPUS addresses cross-modality appearance variation and dataset-specific label spaces through modality-sensitive input adaptation, joint representation learning, prompt-conditioned prediction, and topology-informed deep supervision (Fig. 1).

2.1. Graph-Based Multi-Dataset Backbone

The graph-based multi-dataset backbone comprises modality-specific stems and a modality-shared graph-based encoder-decoder. Let $\mathcal{D} = \{\mathcal{D}^{(d)}\}_{d=1}^K$ denote K datasets, where dataset d provides an

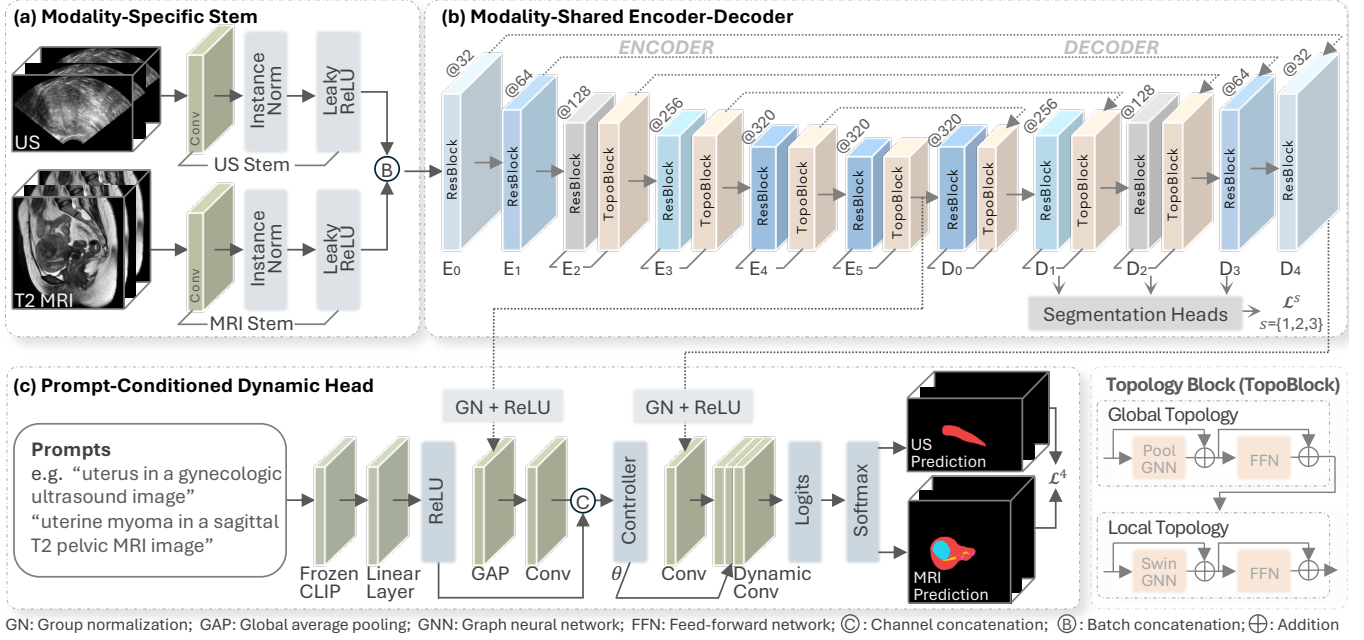


Fig. 1. Overview of the proposed TPUS framework. (a) Modality-specific stems process input 3D US or MRI volumes. (b) The modality-shared graph-based encoder-decoder with topology blocks (TopoBlocks) extracts multi-scale features, with auxiliary heads at decoders D_1 – D_3 . (c) Task-aware class prompts from CLIP [3] and the E_5 bottleneck condition a dynamic head at D_4 for full-resolution prediction.

image $\mathbf{x}^{(d)}$ and a label map $\mathbf{y}^{(d)}$ defined over its own class set $\mathcal{C}_d$. TPUS follows the MultiTalent multi-dataset paradigm [4], while separating modality-sensitive input processing from shared anatomical representation learning. As shown in Fig. 1(a), each dataset has an independent stem $\mathcal{S}_d$ comprising a $3 \times 3 \times 3$ convolution, instance normalization, and LeakyReLU that maps its one-channel volume to 32 channels:

$$\mathbf{z}_0^{(d)} = \mathcal{S}_d(\mathbf{x}^{(d)}). \quad (1)$$

At each iteration, one sampled volume is routed through its corresponding stem. With the reported batch size of one, the concatenation operator in Fig. 1(a) simply forwards the selected stem output rather than pairing US and MRI volumes. The output then enters the modality-shared six-level residual U-shaped encoder-decoder in Fig. 1(b), whose encoder widths are 32, 64, 128, 256, 320, and 320 channels and whose decoder widths are 320, 256, 128, 64, and 32 channels. Resolution-matched skip connections link the two paths; all parameters after the stems are shared.

Within the modality-shared graph-based encoder-decoder, we adapt the Pool graph neural network (GNN) and Swin GNN [5] into several topology blocks (TopoBlocks) deployed at encoder levels E_2 – E_5 and decoder levels D_0 – D_2 ; high-resolution levels E_0 – E_1 and D_3 – D_4 retain convolutional residual blocks. The TopoBlock shown in the bottom right of Fig. 1 applies graph reasoning and feed-forward network (FFN) refinement residually to feature $\mathbf{F}$:

$$\begin{aligned} \hat{\mathbf{F}}_p &= \mathcal{G}_p(\mathbf{F}) + \mathbf{F}, & \mathbf{F}_p &= \text{FFN}(\hat{\mathbf{F}}_p) + \hat{\mathbf{F}}_p, \\ \hat{\mathbf{F}}_w &= \mathcal{G}_w(\mathbf{F}_p) + \mathbf{F}_p, & \mathbf{F}' &= \text{FFN}(\hat{\mathbf{F}}_w) + \hat{\mathbf{F}}_w, \end{aligned} \quad (2)$$

where $\mathcal{G}_p$ denotes Pool GNN, which builds a dynamic k -nearest-neighbor graph after max pooling and restores resolution by max unpooling, and $\mathcal{G}_w$ denotes Swin GNN, which builds dynamic graphs within shifted 3D windows. Both use max-relative graph convolution to complement the backbone’s local convolutional bias.

2.2. Prompt-Conditioned Dynamic Head

Figure 1(c) shows the class-text, image-context, and prediction paths of the dynamic head. For each class $c \in \mathcal{C}_d$, including background, we form a task-aware prompt, such as “uterus in a gynecologic ultrasound image” or “uterine myoma in a sagittal T2 pelvic MRI image.” A frozen CLIP ViT-B/32 text encoder [3] yields an embedding $\mathbf{e}_{d,c} \in \mathbb{R}^{512}$. Then TPUS combines class semantics with image context to generate the classifier. The E_5 bottleneck feature $\mathbf{B}$ and class embedding are projected to 256 dimensions:

$$\begin{aligned} \mathbf{i} &= \text{Conv}(\text{ReLU}(\text{GN}(\mathbf{B}))), \\ \mathbf{t}_{d,c} &= \text{ReLU}(\mathbf{W}_t \mathbf{e}_{d,c} + \mathbf{b}_t), \\ \boldsymbol{\theta}_{d,c} &= \mathcal{H}([\mathbf{i} \parallel \mathbf{t}_{d,c}]), \end{aligned} \quad (3)$$

where $[\cdot \parallel \cdot]$ denotes channel-wise concatenation at the circled C in Fig. 1(c), the shared $1 \times 1 \times 1$ controller $\mathcal{H}$ maps the 512-D condition to $\boldsymbol{\theta}_{d,c} \in \mathbb{R}^{153}$, GAP denotes global average pooling, ReLU represents the rectified linear unit, and GN represents group normalization.

In the prediction path, the final D_4 decoder feature $\mathbf{F}_4 \in \mathbb{R}^{32 \times D \times H \times W}$ is transformed into an eight-channel map:

$$\mathbf{Q} = \text{Conv}(\text{ReLU}(\text{GN}(\mathbf{F}_4))).$$

The dynamic convolution (DynConv) block in Fig. 1(c) comprises three class-wise dynamic convolutions parameterized by $\boldsymbol{\theta}_{d,c}$:

$$\begin{aligned} \mathbf{q}_{d,c}^1 &= \sigma(\mathbf{W}_{d,c}^1 * \mathbf{Q} + \mathbf{b}_{d,c}^1), \\ \mathbf{q}_{d,c}^2 &= \sigma(\mathbf{W}_{d,c}^2 * \mathbf{q}_{d,c}^1 + \mathbf{b}_{d,c}^2), \\ \mathbf{o}_{d,c} &= \mathbf{W}_{d,c}^3 * \mathbf{q}_{d,c}^2 + \mathbf{b}_{d,c}^3, \end{aligned} \quad (4)$$

where $*$ is a $1 \times 1 \times 1$ convolution, σ is ReLU, and the channel dimensions are $8 \rightarrow 8 \rightarrow 8 \rightarrow 1$. Each class therefore requires

$(8^2+8)+(8^2+8)+(8+1) = 153$ generated parameters. The controller is shared across datasets and classes, but the resulting filters depend on both image and class. CLIP embeddings remain fixed, whereas their projection and the controller are trainable. This head replaces the fixed D_4 classifier and produces the native multiclass logits; softmax yields the corresponding prediction. Unlike previous CLIP-driven dynamic convolution [6], which predicts a fixed foreground vocabulary using independent sigmoid outputs, TPUS uses task-aware class prompts to generate background-inclusive softmax logits in each dataset’s native label space.

2.3. Topology-Informed Deep Supervision

We apply deep supervision to outputs at decoder D_s , $s \in \{1, 2, 3, 4\}$. Let $\mathbf{o}^s$ and $\mathbf{y}^s$ denote the logits and corresponding target at D_s . We define pairwise exclusion relations among a dataset’s foreground labels. Unlike previous topology-informed losses [5], which require manually specified relations for a fixed label space, TPUS automatically derives pairwise exclusions from each dataset’s native foreground labels. Hard predictions $\mathbf{P}^s$ are obtained by taking the class-wise argmax of the softmax probabilities. A $3 \times 3 \times 3$ kernel is then applied to $\mathbf{P}^s$ to identify violations within each voxel’s 26-connected neighborhood, yielding a binary critical-voxel mask $\mathbf{V}^s$. For batch size N and voxel domain Ω_s , the topological interaction (TI) term applies cross-entropy only at these voxels:

$$\mathcal{L}_T^s = \frac{1}{N} \sum_{n=1}^N \sum_{v \in \Omega_s} V_{n,v}^s \text{CE}(\mathbf{o}_{n,v}^s, \mathbf{y}_{n,v}^s). \quad (5)$$

For a single foreground class, the exclusion set is empty and $\mathcal{L}_T^s = 0$; the four UMD foreground classes yield six exclusion pairs. The total loss is formulated as follows:

$$\mathcal{L} = \sum_{s=1}^4 w_s \mathcal{L}^s = \sum_{s=1}^4 w_s (\mathcal{L}_{\text{Dice}}^s + \mathcal{L}_{\text{CE}}^s + \lambda_T \mathcal{L}_T^s), \quad (6)$$

where λ_T is empirically set as 10^{-6} . We use normalized deep-supervision weights $w_s = 2^{s-1}/15$ for D_s ($s \in \{1, 2, 3, 4\}$). Only the full-resolution dynamic output at D_4 is used at inference.

3. EXPERIMENTS

3.1. Data and Implementation

The UterUS dataset [1] comprises 3D US volumes with binary uterine-region masks, excluding cases with fibroids, adenomyosis, and other acquired abnormalities. The Uterine Myoma MRI Dataset (UMD) [2] provides labels for the *uterine wall*, *uterine cavity*, *myoma*, and *Nabothian cyst* (when present) on T2-weighted MRI of patients with surgically and pathologically confirmed fibroids. Training/validation/hold-out test sets contain 100/10/28 cases for UterUS and 191/21/54 for UMD. Following MultiTalent [4], we adopt its nnU-Net-based preprocessing [7] (1-mm isotropic resampling and z-score normalization) and data augmentation. Training uses 192^3 patches, batch size 1, and 1,000 epochs of 250 iterations with Nesterov SGD (momentum 0.99, weight decay 3×10^{-5} ; polynomial decay from 10^{-3}). All methods share splits, preprocessing, and postprocessing. Dice and average symmetric surface distance (ASSD) are reported as mean $\pm$ SD over cases per foreground class, then macro-averaged; ASSD is omitted for empty surfaces.

3.2. Experimental Results

We compare the proposed TPUS with task-specific UNETR [8] and SwinUNETR [9], and five joint-training methods: CLIP-UM with

Table 1. Segmentation results on UterUS and UMD.

Method	UterUS		UMD	
	Dice $\uparrow$	ASSD (mm) $\downarrow$	Dice $\uparrow$	ASSD (mm) $\downarrow$
UNETR [8]	0.792 $\pm$ 0.110	0.846* $\pm$ 0.420	0.330 $\pm$ 0.199	18.899* $\pm$ 21.245
SwinUNETR [9]	0.864 $\pm$ 0.071	0.593* $\pm$ 0.304	0.347 $\pm$ 0.200	15.065* $\pm$ 20.817
CLIP-UM [6]	0.861 $\pm$ 0.054	0.637* $\pm$ 0.251	0.144 $\pm$ 0.137	19.165* $\pm$ 11.165
UniSeg [10]	0.498 $\pm$ 0.160	1.715* $\pm$ 0.513	0.599 $\pm$ 0.242	2.732* $\pm$ 4.737
DoDNet [11]	0.891 $\pm$ 0.050	0.760* $\pm$ 1.169	0.433 $\pm$ 0.197	17.143* $\pm$ 19.048
PaNN [12]	0.780 $\pm$ 0.095	2.242* $\pm$ 2.099	0.273 $\pm$ 0.158	20.466* $\pm$ 9.162
MultiTalent [4]	0.895 $\pm$ 0.051	0.448* $\pm$ 0.162	0.680 $\pm$ 0.257	1.775* $\pm$ 3.880
TPUS (Ours)	0.898$\pm$0.050	0.432$\pm$0.156	0.693$\pm$0.254	1.684$\pm$4.680

task-conditioned filters [11], PaNN with size priors [12], and MultiTalent with dataset-specific heads [4]. Table 1 reports results of the eight methods on UterUS and UMD. Table 1 shows that TPUS achieves the best Dice and ASSD on UterUS (0.898 and 0.432 mm vs. 0.895 and 0.448 mm for MultiTalent) and UMD (0.693 and 1.684 mm vs. 0.680 and 1.775 mm). Two-sided paired Wilcoxon signed-rank tests comparing TPUS with each competing method in terms of ASSD demonstrate that TPUS achieves statistically significantly better performance ($p < 0.05$, marked with *).

Figure 2 qualitatively compares different methods on both datasets. On the challenging 3D US dataset (UterUS), competing methods often exhibit target fragmentation or under-segmentation due to acoustic shadowing and low soft-tissue contrast, whereas TPUS produces compact, anatomically contiguous, and boundary-complete uterine masks. On the multi-class T2-weighted MRI dataset (UMD), TPUS suppresses noisy and distant false positives while preserving the uterine wall and cavity and accurately delineating focal lesions, such as myomas and Nabothian cysts, with sharp boundaries. This highlights TPUS’s ability to maintain structural integrity and adapt to heterogeneous multi-structure label spaces.

Table 2. Ablation results on all 54 test cases from UMD.

Configuration	Dice $\uparrow$	ASSD (mm) $\downarrow$
Base	0.677 $\pm$ 0.258	1.834 $\pm$ 4.608
Base + Prompt	0.678 $\pm$ 0.253	1.735 $\pm$ 4.128
Base + Prompt + TopoBlock	0.680 $\pm$ 0.259	1.718 $\pm$ 3.383
Base + Prompt + DynConv	0.684 $\pm$ 0.255	1.606 $\pm$ 4.787
Base + TI	0.686 $\pm$ 0.255	1.536$\pm$3.237
TPUS (Ours)	0.693$\pm$0.254	1.684 $\pm$ 4.680

3.3. Ablation Study

Table 2 evaluates each TPUS component under a matched protocol on UMD. “Base” denotes the MultiTalent model retrained under our protocol. Adding prompt conditioning (Base+Prompt) yields a slight Dice gain, while TopoBlock integration (Base+Prompt+TopoBlock) improves multi-scale long-range reasoning, and dynamic convolution (Base+Prompt+DynConv) enhances dataset-specific adaptation. Topology-informed deep supervision (Base+TI) reduces boundary errors. Combining all components in TPUS achieves the highest Dice, demonstrating their complementary contributions.

3.4. Influence of Prompt-Encoding Strategy

We evaluate the impact of different prompt-encoding strategies within the TPUS framework, comparing our default CLIP text en-

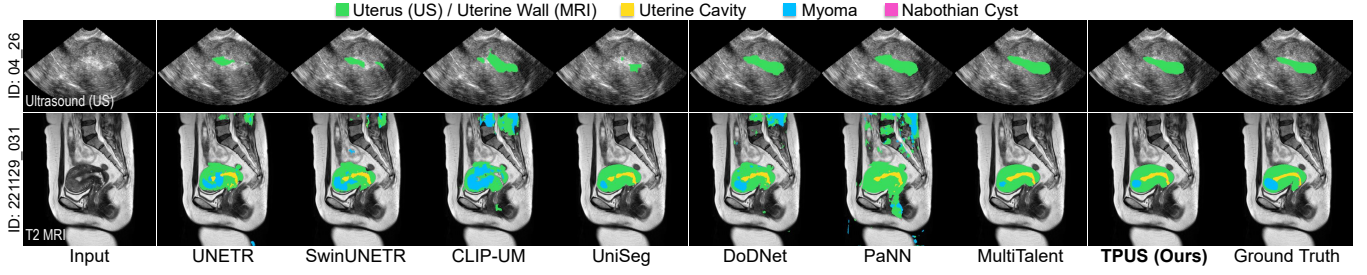


Fig. 2. Visual comparison of segmentation results on the UterUS (top) and UMD (bottom) datasets. Shown are the input, competing methods, the proposed TPUS, and ground truth. Different colors denote distinct tissues.

Table 3. TPUS with different prompt-encoding methods on UterUS.

Method	Dice $\uparrow$	ASSD (mm) $\downarrow$	p -value
Prompt-free FiLM [13]	0.893 ± 0.048	0.456 ± 0.154	1.72×10^{-8}
UniSeg-style Prompt [10]	0.893 ± 0.050	0.463 ± 0.149	1.53×10^{-4}
Random Embedding	0.896 ± 0.047	0.448 ± 0.148	1.17×10^{-3}
CLIP Text Encoder (Ours)	0.898 ± 0.050	0.432 ± 0.156	—

coder against a prompt-free FiLM approach [13], the UniSeg-style task prompt [10], and a randomly initialized learnable embedding. As reported in Table 3, removing explicit prompt conditioning or using alternative schemes slightly degrades segmentation accuracy. The prompt-free FiLM and UniSeg-style prompt variants achieve Dice score of 0.893, while the random embedding strategy reaches 0.896. The proposed pretrained CLIP text encoder achieves the highest Dice and lowest ASSD scores ($p < 0.05$ via two-sided paired Wilcoxon signed-rank tests). These results suggest that semantically rich CLIP-derived prompts used in TPUS provide more effective guidance for segmentation.

4. CONCLUSION

This work presents TPUS, a unified framework for multi-structure uterine segmentation across ultrasound and MRI. TPUS integrates a graph-based multi-dataset backbone comprising modality-specific stems and a modality-shared graph-based encoder-decoder, prompt-conditioned dynamic prediction, and topology-informed deep supervision for heterogeneous uterine ultrasound and MRI datasets. It achieves the highest Dice and lowest ASSD among the compared methods on both held-out test sets. A current limitation is that evaluation is restricted to one ultrasound and one MRI dataset; future work will assess scalability and generalizability by integrating additional datasets with diverse acquisition protocols and label spaces.

5. COMPLIANCE WITH ETHICAL STANDARDS

De-identified UterUS and UMD data were analyzed in this work. The UMD dataset had Beijing Shijitan Hospital ethics approval (sjtkyll-lx-2022(1)) and waived consent; this secondary analysis required no additional approval.

6. ACKNOWLEDGMENTS

This work was supported in part by NIH grants AG073297, AG082938, EB035160, and NS134849. The authors declare no relevant financial or non-financial conflicts of interest.

7. REFERENCES

- [1] E. Boneš, M. Gergolet, C. Bohak, Ž. Lesar, and M. Marolt, “Automatic segmentation and alignment of uterine shapes from 3D ultrasound data,” *Comput. Biol. Med.*, vol. 178, pp. 108794, 2024.
- [2] H. Pan, M. Chen, W. Bai, et al., “Large-scale uterine myoma MRI dataset covering all FIGO types with pixel-level annotations,” *Sci. Data*, vol. 11, no. 1, pp. 410, 2024.
- [3] A. Radford, J. W. Kim, C. Hallacy, et al., “Learning transferable visual models from natural language supervision,” in *ICML*, 2021, vol. 139, pp. 8748–8763.
- [4] C. Ulrich, F. Isensee, T. Wald, M. Zenk, M. Baumgartner, and K. H. Maier-Hein, “MultiTalent: A multi-dataset approach to medical image segmentation,” in *MICCAI*, 2023, pp. 648–658.
- [5] P. Shi, X. Guo, Y. Yang, C. Ye, and T. Ma, “NexToU: Efficient topology-aware U-Net for medical image segmentation,” *arXiv preprint arXiv:2305.15911*, 2023.
- [6] J. Liu, Y. Zhang, J.-N. Chen, et al., “CLIP-driven universal model for organ segmentation and tumor detection,” in *ICCV*, 2023, pp. 21152–21164.
- [7] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [8] A. Hatamizadeh, Y. Tang, V. Nath, et al., “UNETR: Transformers for 3D medical image segmentation,” in *WACV*, 2022, pp. 574–584.
- [9] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, “Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images,” in *BrainLes*, 2022, pp. 272–284.
- [10] Y. Ye, Y. Xie, J. Zhang, Z. Chen, and Y. Xia, “UniSeg: A prompt-driven universal segmentation model as well as a strong representation learner,” in *MICCAI*, 2023, pp. 508–518.
- [11] J. Zhang, Y. Xie, Y. Xia, and C. Shen, “DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets,” in *CVPR*, 2021, pp. 1195–1204.
- [12] Y. Zhou, Z. Li, S. Bai, et al., “Prior-aware neural network for partially-supervised multi-organ segmentation,” in *ICCV*, 2019, pp. 10672–10681.
- [13] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *AAAI*, 2018, pp. 3942–3951.