Neural Radiance Fields for 3D Polyp Segmentation of Deformable Tissues in Endoscopy

Jinhua Liu Wen Tang*
Faculty of Media, Science and Technology, Bournemouth University
Poole BH12 5BB, United Kingdom
jliu2@bournemouth.ac.uk, wtang@bournemouth.ac.uk

Yongsheng Shi Dongjin Huang
Shanghai Film Academy, Shanghai University
Shanghai 200072, China
yongsheng@shu.edu.cn, djhuang@shu.edu.cn

Abstract

Recent advances in Neural Radiance Fields (NeRFs) have demonstrated remarkable segmentation performance in 3D computer vision tasks, capable of segmenting arbitrary content in static scenes. However, it remains a challenge for most NeRFs to perform segmentation in dynamic scenes, such as soft tissue segmentation. In this paper, a novel 3D polyp segmentation method with deformable neural radiance fields is proposed. First, we introduce a new approach called PolypVol-NeRF to extend NeRFs with 3D mask prediction capability. A two-stage training framework is introduced to reconstruct the dynamic radiance field of soft tissue by modelling colour and density. Subsequently, a multi-view self-prompt mechanism is employed to automatically and accurately predict 2D polyp masks for the rendered views generated by the PolypVol-NeRF dynamic implicit field. Finally, we introduce a view-alternating weak supervision strategy to train PolypVol-NeRF for the second time using 2D polyp mask supervision aggregated from multiple training views. This enables obtaining a dynamic 3D Mask implicit field without requiring any ground-truth volumetric annotations. The experimental results demonstrate that our method achieves more comprehensive reconstruction and more accurate segmentation of the lesion region than the state-of-the-art 3D segmentation methods, resulting in the best overall performance. The user study feedback shows that the proposed method has received high ratings from participants in terms of segmentation accuracy, visual quality and clinical applicability.

Keywords: Endoscopy image, Polyp, 3D segmentation, Dynamic neural radiance fields, Image segmentation.

1. Introduction

Reconstructing high-quality deformable tissue from endoscopic images and performing volumetric segmentation of lesions is a challenging task. First, soft tissue deformation and invalid pixel occlusions can interfere with both 3D reconstruction and segmentation. Second, acquiring 3D masks for polyps is extremely labour-intensive and impractical in real surgical scenarios.

With the success of NeRFs in scene reconstruction, a new approach has emerged to project 2D segmentation results of multiple views directly onto 3D meshes or voxel grids for interactive 3D

*Corresponding author: wtang@bournemouth.ac.uk

segmentation within the radiance field. For example, SA3D (Segment Anything in 3D) [5] and SAGA ((Segment Any 3D GAussians) [4] have leveraged powerful visual base models (i.e., Segment Anything Model (SAM)) to accurately segment arbitrary content in 3D static scenes [12]. However, manual interaction for interactive 3D segmentation, although it provides accurate results, is time-consuming and complex for physicians to manage, increasing the operational burden.

Furthermore, most NeRFs variants mainly model the complete static radiance field and lack the capability of differentiating or segmenting individual regions and areas of non-rigid structures. Acquiring 3D masks for polyps is extremely labour-intensive and impractical in real surgical scenarios. As a result, existing methods struggle to achieve accurate 3D segmentation without extensive manual 3D masks. On the other hand, conventional NeRFs excel at reconstructing view-consistent density and color fields. However, they typically lack the capacity to reconcile view-specific 2D semantic predictions into a coherent and spatially aligned 3D representation.

This paper presents a novel 3D polyp segmentation method for accurate 3D polyp segmentation in deformable soft-tissue endoscope scenes, which can also be extended to handle other deformable soft-tissue scenes. The algorithm consists of three modules: a) soft tissue reconstruction; b) multi-view mask generation; c) 2D-to-3D segmentation, as shown in Figure 1.

In the soft tissue reconstruction module, to extract specific polyp regions from the implicit field, we propose a novel method called PolypVol-NeRF. This method adopts a two-stage training framework that enables volumetric mask prediction without ground-truth (GT) volumetric annotations. Aiming to achieve accurate and efficient reconstruction of human soft tissue from monocular endoscopic images, the first stage of PolypVol-NeRF trains only the appearance and geometry components, while keeping the mask branch frozen, to reconstruct the dynamic radiance field.

In the multi-view mask generation module, to address the lack of volumetric mask annotations, we propose a multi-view self-prompt mechanism that combines YOLOv8 and Segment Anything in Medical Images (MedSAM) [15] to generate accurate 2D polyp Masks for all training views quickly. These pseudo masks provide weak semantic cues from different perspectives.

In the 2D-to-3D segmentation module, to lift the multi-view 2D Masks into a coherent 3D segmentation field, we propose a view-alternating weak supervision strategy. This strategy alternately guides the learning of 3D volumetric segmentation in the second-stage PolypVol-NeRF training by leveraging the 2D polyp masks. The trained dynamic 3D Mask implicit field can output precise segmentation of polyps in dynamic scenes.

Experiments on diverse real endoscopic scenes show that our method outperforms existing approaches in both quantitative performance and visual quality. Ablation and user studies further confirm its effectiveness and clinical applicability for reliable 3D polyp segmentation.

2. Related Work

2.1. 3D segmentation

With the development of deep learning, many end-to-end 3D segmentation methods have been proposed based on explicit 3D representations, including volumetric grids, point clouds, and surface meshes [21, 22, 7]. These algorithms perform segmentation directly on pre-existing 3D data and have achieved remarkable success on medical imaging modalities such as CT and MRI. However, these approaches fundamentally rely on explicit 3D inputs and dense volumetric supervision, which are unavailable in clinical endoscopy. Endoscopic images provide only monocular or multi-view 2D observations of deformable tissue surfaces under a limited field of view, while manually annotated 3D lesion ground truth is inherently difficult to obtain in in vivo settings. As a result, meaningful 3D lesion segmentation cannot be directly performed from image observations alone.

For 3D object segmentation from 2D image inputs, some studies treat 3D reconstruction or implicit scene representation as an intermediate

step. Recent NeRFs-based 3D segmentation methods [17] leverage semantic propagation by integrating semantic cues into radiance field representations, enabling label transfer across views in 3D space. Semantic-NeRF [33] and ISRF [8] both explore incorporating semantic cues into NeRF to enable 3D segmentation via semantic propagation. Semantic-NeRF embeds category-level semantics into the radiance field and propagates sparse semantic annotations to novel views for dense multi-class segmentation, whereas ISRF adopts an interactive, category-agnostic setting by propagating sparse user inputs using pretrained 2D visual features without retraining the NeRF model. However, both ISRF and Semantic-NeRF rely heavily on visual priors learned from large-scale natural image datasets and are mainly designed for static or quasi-static scenes. Such methods may not generalise well to domains like endoscopy, where the image characteristics differ significantly from those of natural scenes.

Another line of work explores the integration of 2D segmentation models with NeRF-based 3D understanding [18, 27, 5, 4]. Segment Anything Model (SAM) provides a strong 2D segmentation prior through flexible prompting and strong generalization, making it well suited for supplying reliable 2D cues to downstream 3D segmentation tasks. SAM-Med3D [27] leverages interactive SAM-style prompts to guide volumetric segmentation and achieves accurate and efficient 3D results on 3D medical datasets. Recent studies have explored 3D object segmentation in 3D scene representations reconstructed from 2D images. SA3D [5] enables interactive 3D object selection in NeRF-based scenes using SAM-based prompts from a single view, while SAGA [4] integrates Segment Anything with 3D Gaussian splatting to achieve efficient interactive 3D segmentation. Despite their effectiveness in static scenes, these SAM-based methods rely on manual user prompts and are restricted to static environments, which limits their applicability to dynamic soft-tissue endoscopic scenarios where accurate dynamic soft tissue reconstruction and automatic polyp segmentation from monocular sequences remain challenging.

Unlike conventional 3D segmentation approaches, this paper proposes a new framework for 3D polyp segmentation from monocular 2D endoscopic images. Our method introduces a dedicated multi-view self-prompt mechanism for efficient and accurate lesion localization, and proposes a view-alternating weak supervision strategy that enables hybrid 2D-3D supervision without relying on large-scale manually annotated 3D masks.

2.2. 3D Reconstruction

NeRF is an implicit neural representation that reconstructs 3D scenes from multi-view images, offering high modeling fidelity and superior novel view synthesis quality. With the growing adoption of NeRF and its variants [23, 20], these methods have demonstrated promising potential for 3D reconstruction in endoscopic scenes. Existing studies mainly focus on adapting NeRFs to the specific characteristics of endoscopic data, particularly deformable soft tissues. For instance, EndoNeRF [28] extends D-NeRF [20] to binocular endoscopic sequences and enables detailed reconstruction of deformable surgical scenes. However, such methods rely on paired stereo endoscopic images and are restricted to fixed viewpoints for depth estimation. EndoGaussian [14] and Endo-GSMT [9] extend Gaussian Splatting to monocular endoscopic videos by explicitly modeling dynamic scene elements through learned motion fields and temporal consistency constraints, enabling fast and visually coherent reconstruction under camera motion and soft-tissue deformation. However, these point-based representations are mainly optimized for rendering, while 3D lesion segmentation requires continuous volumetric reasoning and stable topology, which are better supported by implicit representations. Although recent 4D implicit methods focus on dynamic scene reconstruction, their extension to reliable 3D lesion segmentation in endoscopic settings remains largely unexplored.

This paper improves recent 4D implicit representations and proposes a PolypVol-NeRF that enables high-fidelity reconstruction of monocular endoscopic images under challenging condi-

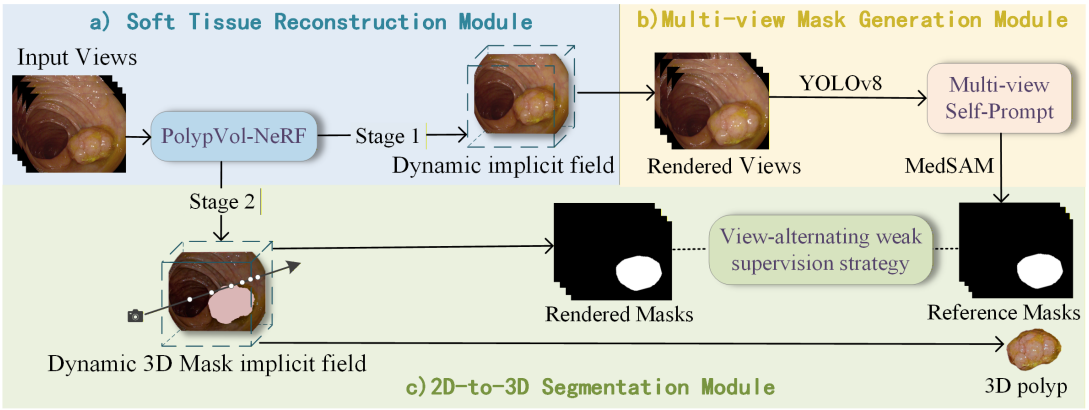

Figure 1: The framework of the proposed 3D polyp segmentation method.

tions, including occlusions caused by invalid pixels and soft-tissue deformations, while simultaneously supporting accurate 3D lesion mask prediction.

3. Methods

Figure 1 shows the framework of our proposed method, which consists of three main modules: soft tissue reconstruction, multi-view mask generation and 2D-to-3D segmentation. The first stage of the proposed PolypVol-NeRF network training yields a dynamic implicit field that enables rendering from arbitrary views for subsequent 2D multi-view mask generation (Sec. 3.1). Next, we use the proposed multi-view self-prompt mechanism to accurately extract the features of rendered views and output 2D Reference Masks that precisely localize the polyps (Sec. 3.2). Finally, our utilizes the proposed view-alternating weak supervision strategy to perform the second stage of training on PolypVol-NeRF. This stage of training enables the mapping of 2D-to-3D polyp mask segmentation (Sec. 3.3).

3.1. Soft Tissue Reconstruction

Existing methods in medical vision rarely address 3D polyp segmentation, especially under the monocular and deformable conditions of real-world endoscopic imaging. Most previous methods rely on simplified geometric representations, which struggle to capture high-frequency details and cannot model view-consistent semantics. We propose a novel semantic-aware framework, PolypVol-NeRF, designed to encode semantic information for 3D object-level segmentation and enable accurate volumetric polyp segmentation. PolypVol-NeRF extends the strengths of Tensor4D [23] in dynamic scene modeling to 3D polyp segmentation, while effectively addressing the unique challenges posed by monocular endoscopic imaging. We first address the issue of invalid pixel occlusion caused by surgical tools and non-informative background regions in endoscopic images by using retrain SAM to predict soft tissue masks that explicitly distinguish valid tissue regions from invalid pixels. Next, the predicted tissue masks are incorporated into Tensor4D through a tissue mask-guided ray sampling strategy, ensuring that the trained dynamic implicit field focus on reconstructing only valid dynamic soft tissue scenes. While these components enable high-fidelity dynamic soft tissue reconstruction, they are insufficient for achieving 3D segmentation. To this end, PolypVol-NeRF introduces an additional mask Multilayer Perception (MLP) network branch that implicitly learns a 3D volumetric mask representation, enabling our framework to incorporate semantic boundary information into the 4D scene representation progressively.

The framework adopts a two-stage training paradigm to separate geometric modeling from semantic learning. This staged design not only stabilizes geometry learning but also effectively decouples appearance modeling from semantic supervision, paving the way for accurate volumetric mask

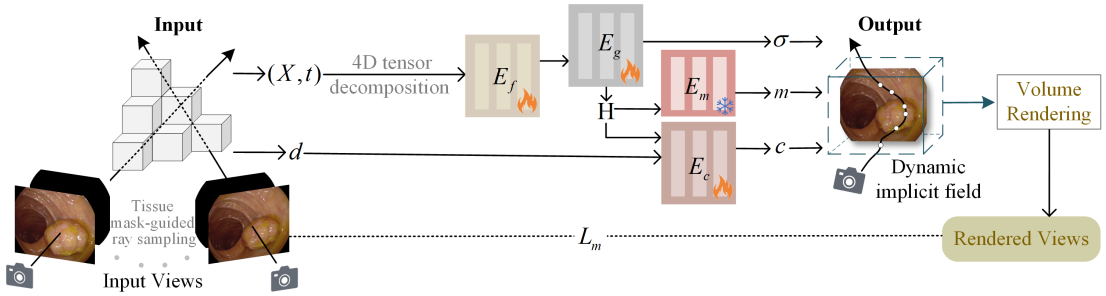

Figure 2: The overall flow of PolypVol-NeRF.

prediction in later stages.

Figure 2 shows the overall flow of PolypVol-NeRF. Given an endoscopic image dataset consisting of 2D input views, PolypVol-NeRF employs a 4D tensor decomposition method to efficiently represent dynamic volumetric scenes. By decomposing the 4D spatiotemporal information (X, t) into time-aware volumes and compact feature planes, it enables the accurate reconstruction of structural motions and dynamic details across multiple scales from coarse to fine [13]. PolypVol-NeRF models a function $f_\phi(X, t, d) \rightarrow (c, \sigma, m)$ through four specialized MLP networks: flow MLP E_f, geometry MLP E_g, color MLP E_c, and mask MLP E_m. This function maps the spatial coordinates $X \in \mathbb{R}^3$, the temporal parameter $t \in \mathbb{R}$ and the view direction d to the corresponding color $c \in \mathbb{R}^3$, volume density $\sigma \in \mathbb{R}$, and volumetric mask $m \in \mathbb{R}$, which are essential for volume rendering and reconstruction. The function f, parameterized by ϕ, is typically represented by a set of MLPs.

Specifically, flow MLP is used to capture the dynamics of the scene in the time dimension, geometry MLP is used to model the geometric structure accurately, color MLP is responsible for predicting the color information of the scene, and the mask MLP decodes semantic mask values for volumetric segmentation. As shown in Figure 2, considering the dependency of semantic prediction on reliable scene geometry, in the soft tissue reconstruction module, we conduct the first-stage training: the mask MLP is entirely frozen, and flow MLP (E_f), geometry MLP (E_g) and color MLP (E_c) are trained to reconstruct view-consistent density and color fields.

To render an image, each pixel is projected into 3D space through ray tracing. A camera ray, denoted as $r(t) = \mathrm{o} + \omega' \cdot \mathrm{d} \cdot t$, originates from the camera's position o and travels in the direction d. The radiance along this ray is accumulated by the volume rendering equation (1) to produce the pixel value $C(r)$:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(p(r(t)))c(p(r(t)), d)dt, \quad (1)$$

where $T(t)$ represents the transmittance from the nearest point t_n to the farthest point t_f, incorporating valid sampling points constrained by the tissue mask. $p(r(t))$ is the 3D point on the camera ray $r(t)$ transformed to the canonical space using MLP network.

During the first stage of network training, PolypVol-NeRF used the loss function (L_m) to optimize the learning of the representation of the volume rendering, ensuring that it is geometrically well structured, feature tight, and color consistent. After optimization of the network training, a dynamic implicit field with a complete representation of the 3D scene is obtained.

3.2. Multi-view Mask Generation

Automated polyp segmentation remains challenging due to the complex appearance and variability of endoscopic images. Many task-specific deep learning models suffer from limited generalization across datasets and imaging conditions. MedSAM is a base model designed for medical image segmentation. It fine-tuned SAM and specifically optimised on a large-scale medical image

dataset of 1,090,486 medical image-mask pairs. The dataset covers 15 imaging modalities and over 30 lesion types to accommodate the specific features and segmentation needs of medical images. Compared with standard polyp segmentation models, MedSAM can meet or even exceed human experts in various medical image segmentation tasks. However, MedSAM requires manual bounding-box input prompts to achieve accurate segmentation and cannot fully automate polyp detection and segmentation.

Achieving accurate 2D polyp segmentation is a fundamental prerequisite, directly influencing the quality of the reconstructed volumetric mask. To automatically and accurately segment polyps in endoscopic images, we propose the multi-view self-prompt mechanism. This mechanism enhances the automation capability of MedSAM in polyp segmentation tasks by combining YOLOv8, thereby improving segmentation accuracy and adaptability. YOLOv8 is capable of quickly and accurately detecting and localising objects within images, while MedSAM has an advantage in segmenting medical soft tissues, diseased areas and tissue structures with high accuracy. The combination of YOLOv8 with MedSAM provides a strong synergy, enabling powerful polyp segmentation. Figure 3 shows the detailed pipeline of the multi-view self-prompt mechanism.

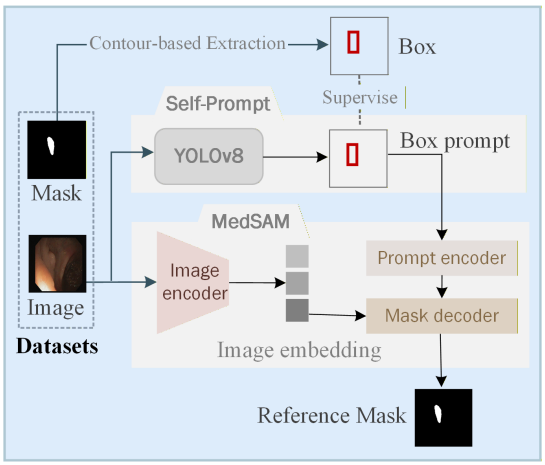

Figure 3: Overall pipeline of the multi-view self-prompt mechanism.

The original YOLOv8 model, trained on natural image datasets, does not generalize well to endoscopic images. To improve its applicability, we retrained YOLOv8 using a task-specific dataset that we constructed for this purpose. We collected 6814 polyp image-mask pairs from the publicly available medical imaging open-source datasets BKAI-IGH NeoPolyp [19], CVC-EndoSceneStill [26], ETIS-LaribPolypDB [24], Kvasir-SEG dataset [10], PolypGen [2] related to gastrointestinal polyps. These image pairs are not directly usable for subsequent polyp detection tasks. To quickly convert 2D Mask information into structured target detection labels. We use the contour-based extraction method to obtain the bounding box of polyps from the collected polyp Masks to realise the conversion from image-mask to image-bounding box. Given image-mask pairs, connected polyp regions Ω_i are first extracted from the binary mask. For each Ω_i, an axis-aligned bounding box B_i is generated as $B_i = (\min x, \min y, \max x, \max y)$, where $(x, y) \in \Omega_i$ denotes pixel coordinates within the polyp region. Small regions are filtered out to suppress noisy annotations. The resulting image-bounding box pairs are formatted in the PASCAL VOC standard and used for detector training. Using this automatically generated dataset, YOLOv8 is retrained while keeping the original network architecture unchanged. By using the trained YOLOv8 for forward inference, bounding boxes for polyp images from any angle can be efficiently obtained. Pre-trained on medical image datasets, MedSAM can accurately predict polyp masks given precise manual bounding-box prompts. Leveraging these complementary strengths, we replace the manual bounding-box prompts with YOLOv8-predicted boxes and feed them, together with the rendered views, directly into MedSAM. This design enables fast and accurate polyp segmentation results (i.e., Reference Mask).

3.3. 2D-to-3D Segmentation

Obtaining high-quality 3D semantic supervision remains difficult in medical imaging, where volumetric annotations are rarely available due to high annotation costs and anatomical complexity. To

generalise the 2D polyp segmentation advantage to segment 3D polyps with dynamic implicit fields, we propose a view-alternating weak supervision strategy that uses pseudo 2D masks generated in Section 3.2 as supervisory signals. These masks are projected back into the 3D space and alternately used to supervise the 3D mask MLP from different viewpoints, enforcing cross-view consistency. This strategy enables the model to transform sparse, view-specific 2D segmentation into a unified, spatially consistent 3D semantic volume.

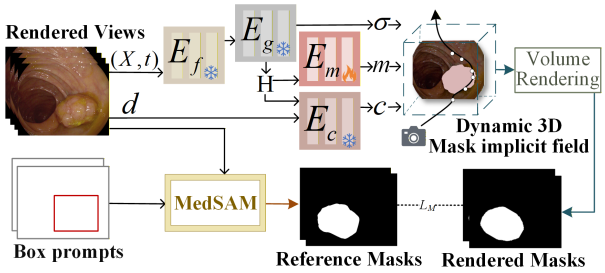

Figure 4: Acquisition process of the dynamic 3D Mask implicit field.

We conducted the second training stage of PolypVol-NeRF to generate a coherent dynamic 3D Mask implicit field. The dynamic 3D Mask implicit field acquisition process is shown in Figure 4. This design aims to obtain 3D polyp Mask of the neural radiance field. 3D Mask can identify and separate voxel data of polyp areas. Only the voxels or areas identified by the 3D Mask will be considered in the light projection. The final generated 3D scene will only contain the target object, while other parts will be excluded or made transparent, thus achieving segmentation of 3D polyps. In this stage, we freeze the flow MLP E_f, geometry MLP E_g and color MLP E_c when training E_m. Feed the high-dimensional feature H into E_m to obtain the Rendered Mask. Then, the Mask Loss L_M is built based on the Rendered Mask and the 2D Reference Mask (GT) outputted by MedSAM, so that the inverse rendering is performed alternately between the training views to iteratively complete the dynamic 3D Mask implicit field training of the target scene. The Mask loss function L_M is formulated as follows:

$$L_M = \frac{1}{N_S} \sum_{i=1}^{N_S} \|M(p,t) - M_{gt}(p,t)\|_2^2 \quad (2)$$

where $M(p,t)$ is the value of the Mask pixel predicted by E_m and $M_{gt}(p,t)$ represents the value of the Rendered Mask pixel.

4. Experiments and Results

4.1. Implementation Details

The experiments were performed in the following computing environment: Windows 10, CPU Intel(R) Xeon(R) E5-2620, CUDA 10.2, Python 3.7, PyTorch 1.5.0, and GPU Nvidia Titan Xp. In our implementation, we trained our model using the Adam optimizer, and set the iteration number to 200K. The batch size of rays was set to 512, and each sampled 64 times along the ray.

This study selected 8 different scenes from the publicly available gastrointestinal medical image datasets, the Gastrolab Image Gallery [1], for qualitative and quantitative analysis. Each scene has the characteristics of occlusion of invalid regions and dynamic changes in soft tissue. All scenes contain 21-65 sequential frame images of size 720 * 576 or 640 * 480. We followed the community standard [17] and held out every 8th image in each scene as a test set for novel view synthesis.

4.2. Qualitative Evaluation

Existing medical 3D segmentation methods cannot be directly applied to 3D lesion segmentation from 2D endoscopic images. Given this limitation, we selected ISRF [8], SA3D [5], and SAGA [4] as comparison baselines, as they represent the closest available methods capable of producing 3D segmentation results in endoscopic scenes from monocular RGB image. Although these methods were originally proposed for general scenes, they provide a challenging and meaningful reference for evaluating 3D segmentation performance. Figure 5 qualitatively compares the segmentation results of the proposed method with those of competing approaches on test Scenes A-D. ISRF produces irregular and scattered regions in Scenes A and C,

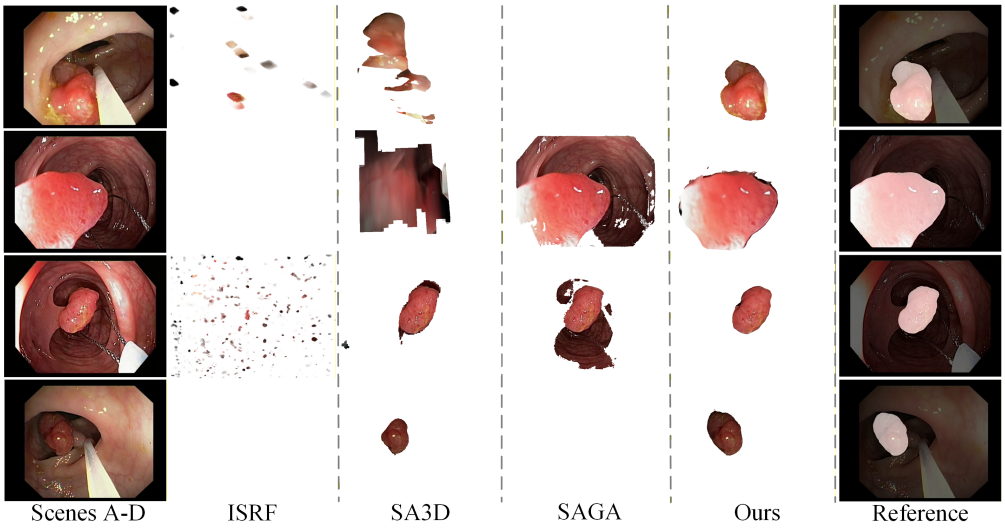

Figure 5: Visual comparison with different 3D segmentation methods for endoscopic images of Scenes A-D.

which deviate significantly from the true lesion areas. SAGA fails to segment any regions in Scenes A and D, and in Scenes B and C it incorrectly labels non-lesion areas as polyps. SA3D performs better than SAGA but still produces inaccurate segmentations and visible artifacts on the reconstructed polyp surfaces. In contrast, the proposed method achieves accurate polyp segmentation while maintaining high-quality scene reconstruction. Overall, our results are the closest to the reference images.

To verify the accuracy of the multi-view self-prompt mechanism proposed in this article for polyp segmentation in endoscopic images, we compared the proposed model with five state-of-the-art polyp segmentation models: ACSNet [32], DCRNet [30], Polyp-PVT [6], NPD-Net [31], and Med-

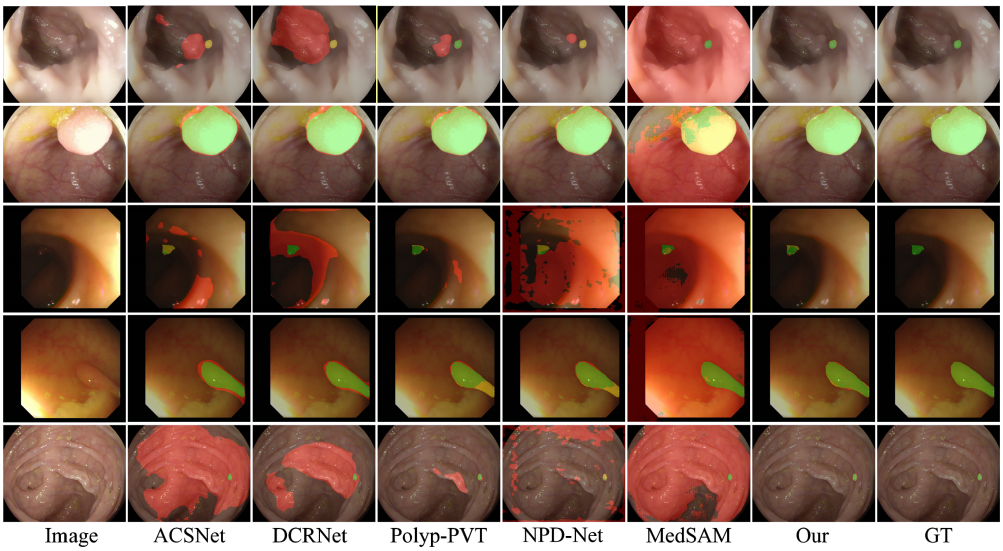

Figure 6: Visual comparison with different polyp segmentation models of the testing set.

SAM [15]. Following the experimental setups in NPD-Net, we selected five challenging public datasets, namely Kvasir SEG [10], ClinicDB [3], ColonDB [25], Endoscene [26] and ETIS [24] to train and test these polyp segmentation models. Figure 6 presents a comparison of the visual effect of different polyp segmentation methods on a diverse set of challenging endoscopic images. These examples cover a wide range of imaging conditions and lesion characteristics, including varying illumination levels, contrast changes, polyp sizes, boundary clarity, and pathological appearances. ACSNet performs poorly on small objects and is sensitive to lighting variations, often leading to over-segmentation or under-segmentation. DCRNet produces imprecise boundaries and tends to over-segment, particularly under poor illumination or for small polyps. Polyp-PVT and NPD-Net both struggle to capture fine-grained details and face challenges in recognizing target regions with ambiguous boundaries. MedSAM's automatic segmentation mode produces large erroneous background regions and suffers from missed segmentation, indicating limited ability to focus on local features in endoscopic images. In contrast, our method robustly segments polyps under complex lighting conditions, performs consistently across different polyp sizes, and achieves high contour overlap even in regions with blurred boundaries. These results suggest that the effective integration of YOLOv8 and MedSAM enables the proposed multi-view self-prompt mechanism to maintain strong robustness and good generalization performance across a wide range of endoscopic imaging conditions.

To validate the scene reconstruction capability of the proposed PolypVol-NeRF, we compare it with the representative 3D reconstruction method EndoNeRF [28] on a binocular endoscopic scene. EndoNeRF relies on binocular image pairs and depth prediction, whereas our method takes only a single-view input from the same binocular sequence for a fair comparison. The visual results are shown in Figure 7. Compared with EndoNeRF, our method better preserves continuous and anatomically plausible vascular-like patterns on the mucosal surface, demonstrating a stronger ability to model fine-grained appearance variations and high-frequency textures. Moreover, in regions affected by surgical instrument occlusions, our method reconstructs more coherent and visually consistent structures, indicating improved robustness to occlusions and incomplete observations. These advantages provide a more reliable basis for subsequent 3D lesion segmentation.

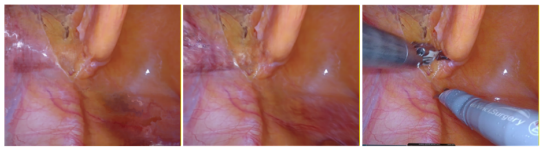

Figure 7: Visual comparison with state-of-the-art 3D reconstruction method for endoscopic images.

4.3. Quantitative Evaluation

We evaluate the proposed automatic 3D polyp segmentation method using three standard metrics: mIoU (mean Intersection over Union) [11], pixel-wise mACC (mean Accuracy) [29], and mDic (mean Dice score) [16]. Table 1 reports the quantitative comparison with ISRF, SA3D, and SAGA across all test scenes. As shown in Table 1, ISRF exhibits the poorest performance, reflecting its limited capability in segmenting polyp regions in endoscopic scenes. SA3D performs slightly better than SAGA, which is consistent with the qualitative results, but still falls behind our method, particularly in mIoU and mDice. Overall, the proposed method consistently outperforms all competing approaches across multiple metrics and scenes. Compared with the best-performing baseline, our method achieves improvements of 81.24% in mIoU, 8.92% in mACC, and 33.40% in mDice, demonstrating significant improvements in segmentation accuracy, overall accuracy, and similarity. These results indicate that our method delivers robust and scalable 3D polyp segmentation performance under diverse endoscopic conditions.

To evaluate the robustness and generalization capability of the proposed multi-view self-prompt mechanism under diverse clinical polyp appear-

Table 1: Quantitative results on Scenes A-H.

Metrics	mIou(%)↑				mACC(%)↑				mDic(%)↑			
Scenes	ISRF	SA3D	SAGA	Ours	ISRF	SA3D	SAGA	Ours	ISRF	SA3D	SAGA	Ours
Scene A	11.83	45.51	40.89	**90.79**	87.39	92.46	93.55	**98.91**	21.12	56.83	54.02	**95.12**
Scene B	0	46.75	39.45	**89.31**	72.32	75.37	80.90	**96.70**	0	63.60	81.96	**94.32**
Scene C	5.70	70.16	35.66	**87.54**	75.99	96.60	87.39	**98.79**	10.78	81.33	51.75	**92.60**
Scene D	0	53.61	4.54	**81.52**	94.29	**96.13**	93.52	95.33	0	69.05	8.70	**89.80**
Scene E	0	70.63	71.77	**92.71**	75.87	92.77	93.14	**99.55**	0	82.67	83.55	**96.20**
Scene F	0	71.45	56.78	**93.87**	80.42	91.91	83.04	**98.76**	0	83.13	69.84	**96.83**
Scene G	0	55.64	58.17	**82.83**	78.29	82.11	88.70	**95.86**	0	71.00	72.70	**90.57**
Scene H	0	35.01	80.73	**84.62**	87.21	71.57	93.32	**97.68**	0	50.87	89.00	**91.53**
Mean	2.19	56.10	48.50	**87.90**	81.47	87.37	89.70	**97.70**	3.99	69.81	63.94	**93.12**

ances and imaging conditions, we conducted a quantitative comparison with five representative 2D polyp segmentation methods. Following common practice, we report six widely used evaluation metrics, including mDic, mIoU, mean absolute error (MAE), weighted F-measure (F_β^ω), S-measure (s_α), and E-measure (E_ξ). For each method, the average performance across all test datasets is reported in Table 2. As shown in Table 2, the original MedSAM consistently under performs compared to other approaches across all metrics. This performance degradation is mainly attributed to the difficulty of accurately localizing polyps using generic prompts. In contrast, the proposed multi-view self-prompt mechanism achieves the best performance across all six metrics, significantly outperforming both MedSAM and other state-of-the-art polyp segmentation methods. This shows that the proposed multi-view self-prompt mechanism generalizes well across diverse clinical polyp appearances and imaging conditions. Moreover, the substantial performance improvement over standalone MedSAM suggests that YOLOv6 maintains strong detection generalization across diverse endoscopic scenes, while MedSAM preserves robust segmentation generalization, enabling reliable lesion localization and mask prediction in our setting.

4.4. Ablation Study

To evaluate the effectiveness of PolypVol-NeRF (Sec. 3.1) and the multi-view self-prompt mechanism (Sec. 3.2), we conduct ablation studies on Scenes A and C. Qualitative results are shown in Figure 8. Here, "w/o Mask MLP" denotes removing the mask MLP branch from PolypVol-NeRF. "w/o Self-Prompt, with auto" correspond to using the original MedSAM in its automatic segmentation mode. "w/o Self-Prompt, with man" denotes the use of the original MedSAM in its manual segmentation mode(bounding-box prompted). In these two settings, the proposed multi-view self-prompt mechanism was removed. As shown in Figure 8, removing the mask MLP (w/o Mask MLP) preserves appearance reconstruction but yields no valid 3D segmentation, since semantic fields cannot be learned without mask supervision. In contrast,

Table 2: Quantitative comparison of different polyp segmentation models on the test set.

Methods	mDic ↑	mIoU ↑	F_β^ω ↑	s_α ↑	E_ξ ↑	MAE ↓
ACSNet	0.8184	0.7492	0.7910	0.8844	0.9034	0.0294
DCRNet	0.7954	0.7330	0.7766	0.8742	0.8924	0.0342
Polyp-PVT	0.8698	0.8038	0.8552	0.9090	0.9466	0.0160
NPD-Net	0.7982	0.7436	0.7892	0.8754	0.8904	0.0180
MedSAM	0.1454	0.0898	0.0920	0.1788	0.1718	0.7514
Our	**0.8758**	**0.8150**	**0.8812**	**0.9132**	**0.9492**	**0.0156**

the improved MedSAM with the proposed multi-view self-prompt mechanism achieves consistently better 3D segmentation than both the automatic and manual MedSAM. The automatic mode produces the weakest results, while the manual mode suffers from user-dependent variability and imprecise ROI localization. These results indicate that 2D mask quality directly affects downstream 3D segmentation, and confirm that the proposed self-prompt mechanism improves segmentation accuracy and consistency by providing more reliable 2D pseudo-masks.

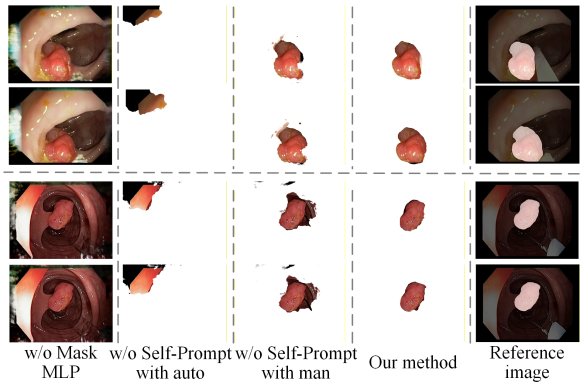

Figure 8: Ablation experiment results on the multi-view self-prompt mechanism. w/o means without.

The quantitative results of the ablation experiments are presented in Table 3, with mIoU, pixel-wise mACC, and mDic used as evaluation metrics. When the mask MLP was removed, the performance dropped substantially across all metrics. Using the MedSAM without the proposed multi-view self-prompt mechanism and operating in the automatic segmentation mode leads to a noticeable performance drop, as this mode produces low-quality polyp masks that are insufficient to effectively guide the learning of 3D segmentation masks. The manual segmentation mode of MedSAM led to moderately improved scores compared to the removal of mask MLP alone, yet still lagged behind the full model. Overall, these results validate that both modules are complementary: mask MLP enhances semantic encoding, while self-prompt further improves the spatial structure of the segmentation.

4.5. User study

We conducted a user study involving 10 participants, including experienced clinicians and senior students with formal training in endoscopy. The participants represented diverse genders and levels of professional experience. The results of ISRF, SA3D, SAGA, and the proposed method were evaluated on 3 randomly selected test images. Each evaluation session lasted at least 20 minutes and was conducted in a quiet and comfortable environment. Participants rated each method using a Likert scale based on segmentation accuracy, visual quality of the segmented polyps, and clinical applicability. A total of 10 valid feedback ratings were collected. In addition, we also received their opinions and suggestions on some open-ended questions.

Figure 9 shows the average scores of different participants considering different aspects on different test images. ISRF consistently received the lowest scores (around 1.8) in all categories, indicating its limited suitability for endoscopic image analysis. In contrast, our method achieved the highest scores in segmentation accuracy and visual quality, demonstrating its superior performance. Participants also rated our method highest in clinical applicability, noting that the generated 3D segmentation results are clear, interpretable,

Table 3: Objective index results of the ablation experiment on Scenes A and C. Values reported in %.

Setting		w/o Mask MLP	w/o Self-Prompt, with auto	w/o Self-Prompt, with man	Ours
Scene A	mIoU ↑	11.02	0.00	83.55	**90.79**
	mACC ↑	11.02	80.74	97.97	**98.91**
	mDic ↑	19.85	0.00	91.03	**95.12**
Scene C	mIoU ↑	6.52	0.00	53.69	**87.54**
	mACC ↑	6.52	83.63	94.68	**98.79**
	mDic ↑	12.25	0.00	69.80	**92.60**

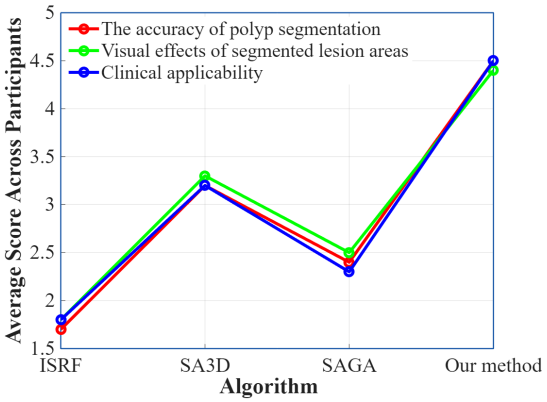

Figure 9: Results of user study.

and beneficial for improving diagnostic accuracy and procedural efficiency. Several clinicians further feedbacked that, although the proposed two-stage pipeline incurs additional computational cost, its offline processing design makes it suited for clinical scenarios such as preoperative planning, retrospective diagnosis, and surgical training, where reconstruction fidelity and structural accuracy are prioritized over real-time performance. Some participants also suggested that closer collaboration with clinicians could further enhance usability, for example through more intuitive user interface design.

5. Discussion

5.1. Discussion on Joint Optimization

A natural question is whether the appearance, geometry, and lesion segmentation can be jointly optimized in an end-to-end manner. While such joint optimization has shown promise in scenarios where reliable 3D supervision is available, it is not well suited to monocular endoscopic settings.

Unlike conventional 3D reconstruction, endoscopic data are limited to monocular image sequences, and reliable 3D lesion annotations are impractical to obtain. Without accurate 3D supervision, joint optimization of geometry and segmentation from the outset can tightly couple inaccurate geometry with noisy 2D pseudo-masks, leading to unstable training and degraded reconstruction quality. To address this challenge, we adopt a two-stage optimization strategy. In the first stage, the model focuses exclusively on learning a geometrically consistent 3D representation of the endoscopic scene from monocular images, supervised only by photometric consistency between rendered views and the input images. This stage establishes a stable and coherent geometry that serves as a prerequisite for reliable 3D segmentation. In the second stage, 2D lesion masks extracted from NeRF-rendered views are introduced as pseudo ground truth to guide the transformation from 2D segmentation to 3D lesion labeling.

This decoupled design enables reliable geometry reconstruction before introducing segmentation supervision, preventing noisy or inconsistent 2D masks from corrupting the 3D representation. In contrast, fully joint optimization of appearance, geometry, and mask would require accurate 3D supervision or strong multi-view constraints, which are unavailable in monocular endoscopic settings. Our experiments further show that 2D segmentation quality directly affects 3D performance, underscoring the need to first establish a robust geometric foundation.

5.2. Limitations and Future Work

One limitation of the proposed framework is its reliance on a two-stage pipeline with multiple submodules, which introduces additional computational complexity and training overhead. However, this design enables accurate 3D lesion segmentation that is difficult to achieve with real-time or single-stage approaches, particularly in endoscopic scenarios where explicit 3D supervision is unavailable. Future work will explore more efficient scene representations and streamlined optimization strategies to reduce computational cost without compromising 3D segmentation accuracy, thereby improving training efficiency and facilitating more practical deployment.

6. Conclusion

We presented a novel framework for 3D polyp segmentation that integrates three key components: soft tissue reconstruction, multi-view mask generation, and 2D-to-3D segmentation. These modules are designed to jointly capture appearance, geom-

etry, and semantics in a unified 3D representation. The training process is divided into two stages. In the first stage, the model learns to reconstruct the surgical scene from monocular endoscopic images, leveraging appearance and geometry cues via a radiance field representation. In the second stage, the mask MLP is optimized to infer volumetric segmentation masks using sparse 2D mask supervision. Experiments show our method achieves more accurate 3D segmentation results than state-of-the-art 3D segmentation methods, helping doctors quickly and clearly observe the lesion site.

Acknowledgement

This work was supported by the Marie Skłodowska-Curie Actions Postdoctoral Fellowships under the European Union HORIZON-MSCA-2024-PF-01 programme (Grant No. 101202418).

Conflict of Interest

The authors declare no conflicts of interest.

References

[1] Gastrolab. The gastrolab image gallery. http://www.gastrolab.net/index.htm. Accessed: 2024-11-11. 7

[2] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, M. A. Riegler, K. V. Anonsen, et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10(1):75, 2023. 6

[3] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 9

[4] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian. Segment any 3d gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1971–1979, 2025. 2, 3, 7

[5] J. Cen, Z. Zhou, J. Fang, W. Shen, L. Xie, D. Jiang, X. Zhang, Q. Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36:25971–25990, 2023. 2, 3, 7

[6] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021. 8

[7] Y. Du, F. Bai, T. Huang, and B. Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 37:110746–110783, 2024. 2

[8] R. Goel, D. Sirikonda, S. Saini, and P. Narayanan. Interactive segmentation of radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4201–4211, 2023. 3, 7

[9] H. Gou, C. Wang, J. Yang, Y. Liu, F. Jia, D. Xiao, F. Qin, and H. Luo. Endo-gsmt: Endoscopic monocular scene reconstruction with dynamic gaussian splatting and motion tracking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 213–223. Springer, 2025. 3

[10] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 6, 9

[11] Z. Ke, Y. Wang, R. Guo, M. Du, J.-S. Zhou, G. Wang, and Y. Zhang. An effective algorithm for skin disease segmentation combining inter-channel features and spatial feature enhancement. In *International Conference on Computational Visual Media*, pages 109–128. Springer, 2025. 9

[12] 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*, pages 4015–4026, 2023. 2

[13] J. Liu, Y. Shi, D. Huang, and J. Qu. Neural radiance fields for high-fidelity soft tissue reconstruction in endoscopy. *Sensors*, 25(2):565, 2025. 5

[14] Y. Liu, C. Li, C. Yang, and Y. Yuan. Endogaussian: Real-time gaussian splatting for dynamic endoscopic scene reconstruction. *arXiv preprint arXiv:2401.12561*, 2024. 3

[15] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2, 9

[16] W. Meng, X. Zhu, and Y. Li. Ynet: Medical image segmentation model based on wavelet trans-

form boundary enhancement. In *International Conference on Computational Visual Media*, pages 91–108. Springer, 2025. 9

[17] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. 3, 7

[18] A. Mirzaei, T. Aumentado-Armstrong, K. G. Derpanis, J. Kelly, M. A. Brubaker, I. Gilitschenski, and A. Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. 3

[19] P. Ngoc Lan, N. S. An, D. V. Hang, D. V. Long, T. Q. Trung, N. T. Thuy, and D. V. Sang. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In *International Symposium on Visual Computing*, pages 15–28. Springer, 2021. 6

[20] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. 3

[21] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[22] L. Schneider, A. Niemann, O. Beuing, B. Preim, and S. Saalfeld. Medmeshcnn-enabling meshcnn for medical surface models. *Computer Methods and Programs in Biomedicine*, 210:106372, 2021. 2

[23] R. Shao, Z. Zheng, H. Tu, B. Liu, H. Zhang, and Y. Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 3, 4

[24] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014. 6, 9

[25] N. Tajbakhsh, S. R. Gurudu, and J. Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 9

[26] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017. 6, 9

[27] H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, J. Chen, Y. Su, Z. Huang, Y. Shen, et al. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. In *European Conference on Computer Vision*, pages 51–67. Springer, 2024. 3

[28] Y. Wang, Y. Long, S. H. Fan, and Q. Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International conference on medical image computing and computer-assisted intervention*, pages 431–441. Springer, 2022. 3, 9

[29] Z. Yang, H. Wang, and F. Zhang. Hifnet: Medical image segmentation network utilizing hierarchical attention feature fusion. In *International Conference on Computational Visual Media*, pages 74–90. Springer, 2025. 9

[30] Z. Yin, K. Liang, Z. Ma, and J. Guo. Duplex contextual relation network for polyp segmentation. In *2022 IEEE 19th international symposium on biomedical imaging (ISBI)*, pages 1–5. IEEE, 2022. 8

[31] Z. Yu, L. Zhao, T. Liao, X. Zhang, G. Chen, and G. Xiao. A novel non-pretrained deep supervision network for polyp segmentation. *Pattern Recognition*, 154:110554, 2024. 8

[32] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu. Adaptive context selection for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 253–262. Springer, 2020. 8

[33] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 3