High-fidelity Endoscopic Image Synthesis by Utilizing Depth-guided Neural Surfaces

Baoru Huang[1,2*], Yida Wang[3,*], Anh Nguyen[4], Daniel Elson[1], Francisco Vasconcelos[2], Danail Stoyanov[2]

[1]Imperial College London, UK [2]University College London, UK [3]Technical University of Munich, Germany

[4]University of Liverpool, UK *Co-First authors

Abstract

In surgical oncology, screening colonoscopy plays a pivotal role in providing diagnostic assistance, such as biopsy, and facilitating surgical navigation, particularly in polyp detection. Computer-assisted endoscopic surgery has recently gained attention and amalgamated various 3D computer vision techniques, including camera localization, depth estimation, surface reconstruction, etc. Neural Radiance Fields (NeRFs) and Neural Implicit Surfaces (NeuS) have emerged as promising methodologies for deriving accurate 3D surface models from sets of registered images, addressing the limitations of existing colon reconstruction approaches stemming from constrained camera movement.

However, the inadequate tissue texture representation and confused scale problem in monocular colonoscopic image reconstruction still impede the progress of the final rendering results. In this paper, we introduce a novel method for colon section reconstruction by leveraging NeuS applied to endoscopic images, supplemented by a single frame of depth map. Notably, we pioneered the exploration of utilizing only one frame depth map in photorealistic reconstruction and neural rendering applications while this single depth map can be easily obtainable from other monocular depth estimation networks with an object scale. Through rigorous experimentation and validation on phantom imagery, our approach demonstrates exceptional accuracy in completely rendering colon sections, even capturing unseen portions of the surface. This breakthrough opens avenues for achieving stable and consistently scaled reconstructions, promising enhanced quality in cancer screening procedures and treatment interventions.

1. Introduction

Colorectal cancer stands as the third most frequently diagnosed cancer and the second leading cause of cancer-related mortality [18]. Timely detection is paramount for favourable prognoses. While various techniques such as virtual colonoscopy exist, optical colonoscopy remains the gold standard for screening and lesion removal. As computer-assisted systems gain prominence in routine endoscopic procedures and present a fertile ground for innovation, typical computer vision approaches are applied including but not limited to depth estimation, 3D reconstruction, and camera localization, benefiting advancements in polyp detection, stenosis assessment, post-intervention diagnosis, and exploration thoroughness evaluation. Nevertheless, achieving precise 3D reconstructions of large colon sections in endoscopic videos remains challenging due to constrained camera movement and the dearth of texture information within the colon.

To address this challenge, prior research has demonstrated the feasibility of estimating the colon's 3D shape from single images captured during colonoscopies [1]. Yet, achieving dense reconstructions for large sections necessitates the utilization of multiple images. Since most endoscopes are equipped with a monocular camera, leveraging video sequences using structure-from-motion algorithms emerges as a natural approach [9]. Despite advances in image registration techniques [16] and SLAM [3], sparse reconstructions remain prevalent, leaving the transition to dense reconstructions unresolved.

The emergence of Neural Radiance Field (NeRF) [13] networks presents a promising avenue for acquiring implicit 3D representations from image sets. NeRF leverages neural implicit fields for continuous scene representations, demonstrating remarkable success in high-quality view synthesis and 3D reconstruction [12]. Instant NGP [14] reduced computational costs by introducing versatile new input encoding techniques. Recent developments like Gaussian NeRF [8] offer representations using 3D Gaussians, preserving desirable properties of continuous volumetric radiance fields for scene optimization. However, traditional NeRF-based approaches are ill-suited for endoscopic surgery videos. Unlike conventional scenes where cameras capture images from various viewpoints, endoscopic cameras operate within confined spaces, like the cylindrical tunnel of the colon, severely limiting viewing directions

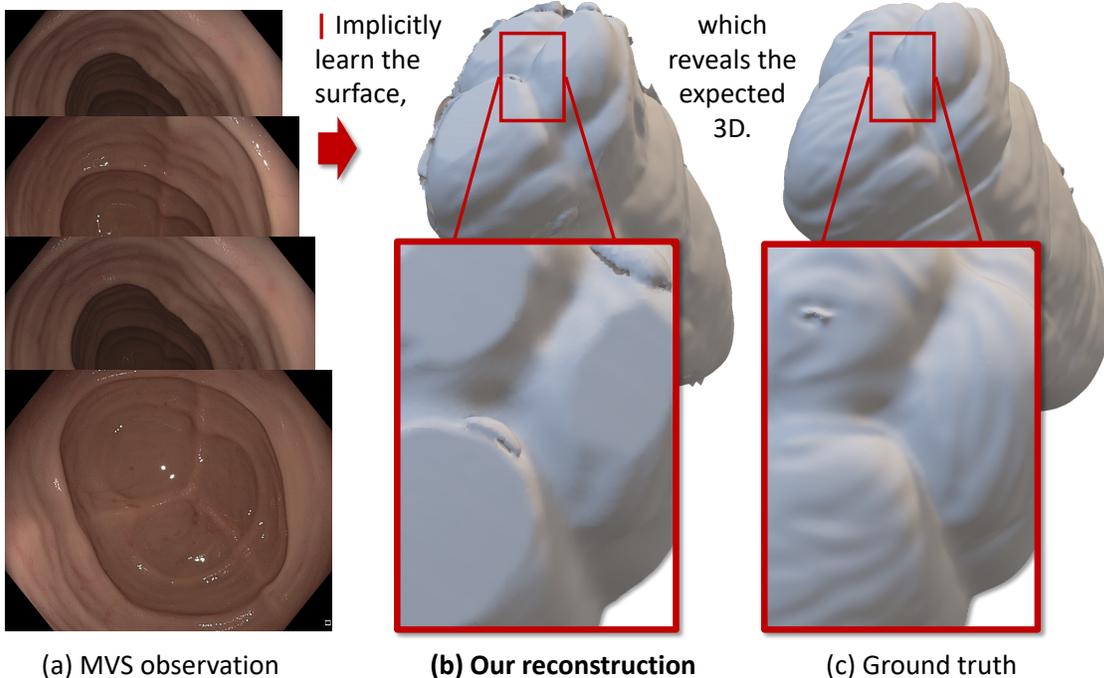| (a) MVS observation | **(b) Our reconstruction** | (c) Ground truth |

Figure 1. Implicitly reconstructing 3D structures (b) from MVS (multi-view stereo) images (a), which precisely matches the ground truth 3D structures in (c).

and camera movement. Consequently, existing methods, including Neural Implicit Surfaces (NeuS) [19], which represent surfaces as zero-level sets of signed distance functions (SDFs), fail to provide consistent depth mapping in endoscopic scenarios. Therefore, there is a critical need for novel approaches tailored specifically to the unique challenges posed by endoscopic environments.

Our paper presents a new approach to incorporating depth information into NeuS for endoscopy reconstruction. Instead of relying on a dedicated dataset for training a depth estimator [5–7] tailored to surgical scenes, we leverage an existing depth estimator trained on nature scenes. By directly inferring depth maps from this pre-trained estimator and rescaling them to the endoscopy scenes, we effectively supervise the NeuS model. The intensive experiments show that our method achieves state-of-the-art results and produces accurate depth maps on 2D synthetic rendering and 3D reconstruction of endoscopic surgical scenes.

2. Methodology

Given a set of calibrated multi-view images that capture an object or a static scene with their corresponding camera poses, a set of the structural surface and the appearance $\{\mathbf{S}, \mathbf{C}\}$ of the targeted environment could be learned through the radiance supervision [11, 19, 20]. The learned set $\{\mathbf{S}, \mathbf{C}\}$ is represented by the signed distance field (SDF) $f(x) : \mathbb{R}^3 \to \mathbb{R}$ where the value of each element is de-

termined by a 3D position x, and a radiance field $c(x, v)$: $\mathbb{R}^3 \times \mathbb{S}^2 \to \mathbb{R}^3$ which is determined by both the position x and the viewing direction $v \in \mathbb{S}^2$. Aiming at learning more precise zero-crossing surfaces in SDF by jointly training the SDF and the radiance field, we introduce two proposed a factor λ_{r} to make the SDF regularization more adaptive to improve the rendering quality and reduce the geometric bias.

2.1. Neural rendering

By enforcing the radiance supervision through the 2D image, NeRF [13] leverages volume rendering to match the ground truth for every camera pose with the rendered image. Specifically, the RGB for every pixel of an image can be generated by sampling n points $\{ r(t_i) = o + t_i \cdot v \mid i = 1, \ldots, n \}$ along its camera ray r, where o is the center of the camera, t_i is the sampling interval along the ray and v is the view direction. By accumulating the radiance field density $\sigma(r(t))$ and the colors $c(r(t), v)$ of the sample points, the color $\hat{\mathbf{C}}$ of the ray could be presented as

$$\hat{\mathbf{C}}(r) = \int_{t_{\mathrm{n}}}^{t_{\mathrm{f}}} T(t) \cdot \sigma(r(t)) \cdot c(r(t), v) \, dt \qquad (1)$$

where the transparency $T(t)$ is derived from the volume density $\sigma(r(t))$. $T(t)$ denotes the accumulated transmittance along the ray r from the closest point t_{n} to the farthest

point t_f such that

$$T(t) = \exp\left(-\int_{t_\text{n}}^{t} \sigma(r(s))ds\right) . \quad (2)$$

Note that $T(t)$ is a monotonic decreasing function with a starting value $T(t_\text{n})$ of 1. The product $T(t) \cdot \sigma(r(t))$ is used as a weight $\omega(t)$ in the volume rendering of the radiance in Eq. (1).

Since the rendering process is differentiable, our model can then learn the radiance field c from the multi-view images with the loss function that minimizes the color difference between the rendered pixels $\hat{\mathbf{C}}(r)$ with $i \in \{1, \ldots, m\}$ and the corresponding ground truth pixels $\mathbf{C}(r)$ without 3D supervision as

$$\mathcal{L}_\text{rgb} = \frac{1}{m}\sum_{i=1}^{m}\left(\|\hat{\mathbf{C}}(r) - \mathbf{C}(r)\|_2 + |\hat{\mathbf{C}}(r) - \mathbf{C}(r)|\right) \quad (3)$$

where m denotes the batch size during training. Based on the same input and output, we would further investigate a way to implicitly learn a signed distance field f to extract meshes embedded in Eq. (1) during training.

2.2. Depth-guided SDF optimization

Aiming at extracting a 3D mesh from a region of interest in the neural rendering, it is plausible to get a projection from a signed distance function (SDF) to the radiance field. Here, we look for a function Φ that transforms the signed distance function so that it can be used to compute the density-related term $T(t)\sigma(r(t))$ in Eq. (1).

We build our solution on top of HF-NeuS [20], where they set $\Phi(r(t))$ as the transparency $T(t)$. Notably, the derivative of the transparency function $T(t)$ is the negative weighting function as

$$\frac{d(T(t))}{dt} = -T(t)\sigma(r(t)) . \quad (4)$$

Given such formulation, the SDF surface lies on the maximum radiance weight. The maxima is computed by setting the derivative of the weighting function to zero; thus,

$$\frac{d(T(t)\sigma(r(t)))}{dt} = -\frac{d^2(T(t))}{dt^2} = -\frac{d(T'(t))}{dt} = 0 . \quad (5)$$

To fulfill the criteria from Eq. (4) and (5), HF-NeuS [20] defined the transparency function $T_\text{s}(t)$ as the normalized sigmoid function $[1 + \exp(s \cdot f(r(t)))]^{-1}$ with a parameter s. The scalar s reveals how strong the SDF is related to the radiance field, which is usually increasing during training. In the initial stage of training, a small s relaxes the connection between SDF and radiance such that the radiance parametric model is optimized without a dependency on the plausible SDF, where the view-conditioned depth $\mathbf{Z}(r_i)$ for each ray r_i matches the ground truth depth $\hat{\mathbf{Z}}(r_i)$.

To parameterize SDF, a signed distance function $f(x)$ is differentiable almost everywhere, while its gradient $\boldsymbol{\nabla}f(x)$ satisfies the Eikonal equation $\|\boldsymbol{\nabla}f(x)\|_2 = 1$. This implies that an SDF can be trained with the Eikonal regularization. According to IGR [4], we enforce an Eikonal loss as a regularizer to make an implicit field act as a signed distance field. However, we discovered that training with a fixed Eikonal regularization [19, 20, 22] can lead to a sub-optimal convergence where the SDF has converged based on the Eikonal regularization but the rendering can still be further optimized. Such a problem harms the improvement of the RGB rendering because the weights are projected from a wrong SDF value, which makes the extracted mesh miss the detailed structures.

To solve the problem of the reconstruction failures on low-textured structures, *i.e.* internal surface of organ, we found that the capability of rendering such structures from the radiance field should be ensured even when the SDF does not contain the accorded structures so that the gradient could be back-propagated from the radiance field to the SDF later on. Given the overall loss function $\mathcal{L}_\text{total} = \mathcal{L}_\text{rgb} + \mathcal{L}_\text{sdf}$, RaNeuS [21] proposed to adaptively weight the Eikonal regularization to optimize the SDF as

$$\mathcal{L}_\text{sdf} = \frac{\lambda_\text{E}}{mn}\sum_{i=1}^{m}\lambda_\text{r}(r_i)\sum_{j=1}^{n}(\|\mathbf{n}_{ij}\|_2 - 1)^2 , \quad (6)$$

where a ray-wise weight $\lambda_\text{r}(r_i)$ for the i-th ray r_i is set to be

$$\lambda_\text{r}(r_i) = \frac{\alpha}{d_\text{r}(r_i) + \alpha} . \quad (7)$$

We propose to present the ray-wise weight $\lambda_\text{r}(r_i)$ considering a depth bias instead. Given the depth value $\mathbf{Z}(r_i)$ determined by the zero-crossing point sampled from ray r_i, $d_\text{r}(r_i)$ is the depth distance $\|\hat{\mathbf{Z}}(r_i) - \mathbf{Z}(r_i)\|_2$ regarding ray r_i, and α is a positive hyperparameter that is set to be smaller than 1, *e.g.* $1 \cdot 10^{-3}$. Notice that a precise per-ray depth supervision $\hat{\mathbf{Z}}(r_i)$ is not necessarily supplemented, while the depth distance $d_\text{r}(r_i)$ is not optimized directly here. Given a general depth estimation model such as a DPT Hybrid model [10], we rescale the predicted depth according to a reconstructed MVS point cloud using COLMAP [15, 17] to represent an approximation of $\hat{\mathbf{Z}}(r_i)$. Practically, the gradients back-propagated to the zero crossing point $\mathbf{Z}(r_i)$ are disabled during training. Consequently, the Eikonal regularization will be relaxed when the geometric precision determined by the metric of $d_\text{r}(r_i)$ is unsatisfying.

We follow the remaining hyper-parameter setups of RaNeuS [21], where the approximated normal $\mathbf{n} = \boldsymbol{\nabla}f(r(\cdot))$ is the derivative of $f(r(\cdot))$, and λ_E is typically set to be 0.1 as mentioned by IGR [4] and NeuS [19].

Method	Item	c1_a	c1_b	c2_a	c2_b	c2_c	c3_a	c4_a	c4_b	d4_a	s1_a	s2_a
Gaussian NeRF [8]	psnr	32.84	28.19	30.24	**31.07**	28.25	26.53	28.41	32.03	29.26	**36.27**	35.22
Vanilla NeRF [13]	psnr	32.71	30.68	31.35	29.32	**32.32**	**33.42**	32.35	33.70	28.65	35.92	31.00
Ours	psnr	**34.27**	**31.02**	**33.78**	31.04	31.02	32.75	**33.62**	**34.19**	**31.01**	31.24	**35.88**

Table 1. Quantitative results. The best result in each small sequence is presented in bold.

To avoid extreme values for different scenes, $d_r(r)$ is set to be zero when a ray does not hit the foreground SDF space. A converged model is respected to end up with the typical Eikonal regularization when there is precise surface geometry measured by a small $d_r(r)$. Thus, in contrast to [4], our approach is more adaptive to the changes in both the rendering and SDF optimizations; thus, relaxing the restrictions from the predefined parameters.

3. Experimental Results

Our method was trained and assessed using C3VD [2], a dataset comprising small video sequences obtained from real wide-angle colonoscopies at a resolution of 1350 × 1080. These sequences traverse four distinct colon phantoms, including the colon cecum, descending, sigmoid, and transcending regions. Video lengths vary from 61 to 1142 frames, with per-frame camera poses utilized. We divided each scene into training, testing, and validation sets in a 6:2:2 ratio. Additionally, we compared our model's performance with that of Gaussian NeRF [8] and vanilla NeRF.

The results, presented in Table 1, were evaluated based on RGB reconstruction quality, measured using peak signal-to-noise ratio (PSNR). Notably, our method consistently outperforms or matches the performance of Gaussian NeRF [8] and vanilla NeRF [13] across most scenarios, as indicated in Table 1.

4. Conclusion

In conclusion, we have introduced an approach to enhance the reconstruction and rendering quality of endoscopic images by integrating single-frame depth-guided SDF optimization. We employed adaptive regularization to mitigate geometric bias, and our method's efficacy was demonstrated through PSNR metrics. Remarkably, even without depth guidance, our method surpassed the performance of Gaussian NeRF [8] and vanilla NeRF [13]. Moreover, the incorporation of depth guidance resulted in further improvements in neural rendering quality.

References

[1] Víctor M Batlle, José MM Montiel, and Juan D Tardós. Photometric single-view dense 3d reconstruction in endoscopy. In *IROS*, 2022. 1

[2] Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S Akshintala, Juan R Garcia, and Nicholas J Durr. Colonoscopy 3d video dataset with paired depth from 2d-3d registration. *Medical image analysis*, 2023. 4

[3] Juan J Gómez-Rodríguez, José Lamarca, Javier Morlana, Juan D Tardós, and José MM Montiel. Sd-defslam: Semidirect monocular slam for deformable and intracorporeal scenes. In *ICRA*, 2021. 1

[4] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 3, 4

[5] Baoru Huang, Jian-Qing Zheng, Anh Nguyen, David Tuch, Kunal Vyas, Stamatia Giannarou, and Daniel S Elson. Self-supervised generative adversarial network for depth estimation in laparoscopic images. In *MICCAI*, 2021. 2

[6] Baoru Huang, Anh Nguyen, Siyao Wang, Ziyang Wang, Erik Mayer, David Tuch, Kunal Vyas, Stamatia Giannarou, and Daniel S Elson. Simultaneous depth estimation and surgical tool segmentation in laparoscopic images. *TMRB*, 2022.

[7] Baoru Huang, Jian-Qing Zheng, Anh Nguyen, Chi Xu, Ioannis Gkouzionis, Kunal Vyas, David Tuch, Stamatia Giannarou, and Daniel S Elson. Self-supervised depth estimation in laparoscopic image using 3d geometric consistency. In *MICCAI*, 2022. 2

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. 1, 4

[9] Simon Leonard, Austin Reiter, Ayushi Sinha, Masaru Ishii, Russell H Taylor, and Gregory D Hager. Image-based navigation for functional endoscopic sinus surgery using structure from motion. In *Medical Imaging*, 2016. 1

[10] Chunpu Liu, Guanglei Yang, Wangmeng Zuo, and Tianyi Zang. Dpdformer: a coarse-to-fine model for monocular depth estimation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 3

[11] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Yuan Liu, Peng Wang, Christian Theobalt, Taku Komura, and Wenping Wang. Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. *arXiv preprint arXiv:2211.14173*, 2022. 2

[12] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 1

[13] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 1, 2, 4

[14] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a mul-

tiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 2022. 1

[15] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3

[16] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[17] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 3

[18] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 2021. 1

[19] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2, 3

[20] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. Hf-neus: Improved surface reconstruction using high-frequency details. *NeurIPS*, 2022. 2, 3

[21] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Raneus: Ray-adaptive neural surface reconstruction. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024. 3

[22] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *ACM Transactions on Graphics (TOG)*, 2022. 3