EndoSelf: Self-Supervised Monocular 3D Scene Reconstruction of Deformable Tissues with Neural Radiance Fields on Endoscopic Videos

Wenda Li[1], Yuichiro Hayashi[1], Masahiro Oda[2,1], Takayuki Kitasaka[3], Kazunari Misawa[4], and Kensaku Mori[1,2,5]

[1] Graduate School of Informatics, Nagoya University, Nagoya, 464-8601, Aichi, Japan
wdli@mori.m.is.nagoya-u.ac.jp
kensaku@is.nagoya-u.ac.jp
[2] Information Technology Center, Nagoya University, Nagoya, 464-8601, Aichi, Japan
[3] Faculty of Information Science, Aichi Institute of Technology, Yakusacho, Toyota, 470-0392, Aichi, Japan
[4] Aichi Cancer Center Hospital, Nagoya, 464-8681, Aichi, Japan
[5] Research Center of Medical Bigdata, National Institute of Informatics, Hitotsubashi, 101-8430, Tokyo, Japan

Abstract. Neural radiance fields have recently emerged as a powerful representation to reconstruct deformable tissues from endoscopic videos. Previous methods mainly focus on depth-supervised approaches based on endoscopic datasets. As additional information, depth values were proven important in reconstructing deformable tissues by previous methods. However, collecting a large number of datasets with accurate depth values limits the applicability of these approaches for endoscopic scenes. To address this issue, we propose a novel self-supervised monocular 3D scene reconstruction method based on neural radiance fields without prior depth as supervision. We consider the monocular 3D reconstruction based on two approaches: ray-tracing-based neural radiance fields and structure-from-motion-based photogrammetry. We introduce structure from motion framework and leverage color values as a supervision to complete the self-supervised learning strategy. In addition, we predict the depth values from neural radiance fields and enforce the geometric constraint for depth values from adjacent views. Moreover, we propose a looped loss function to fully explore the temporal correlation between input images. The experimental results showed that the proposed method without prior depth outperformed the previous depth-supervised methods on two endoscopic datasets. Our code is available at https://github.com/MoriLabNU/EndoSelf.

Keywords: 3D reconstruction · Self-supervised learning · Neural radiance fields.

1 Introduction

3D scene reconstruction has emerged as a crucial field of study within minimally invasive surgery (MIS) [9], offering significant advancements in surgical

procedures. Reconstructing scenes from endoscopic videos provides a broader and more detailed 3D Field of View (FoV) in surgical navigation systems [18]. Moreover, it is used to facilitate the automation of robot-assisted MIS. The 3D models of scenes and organs benefit virtual reality (VR) and augmented reality (AR) for MIS [4, 24]. It allows for preoperative planning and surgical education with efficient creation of surgical scenes [26].

Structure-from-motion-based photogrammetry and ray-tracing-based neural radiance fields perform remarkably in monocular 3D scene reconstruction approaches. Photogrammetry includes two foundational approaches to realize 3D scene reconstruction for different datasets. For stereo images, the approach involves a systematic feature extraction process and the application of epipolar geometry to facilitate accurate feature-matching across stereo views [8]. This process culminates in generating depth maps converted from disparity maps based on the stereo camera's intrinsic parameters. For monocular videos, the approach revolves around estimating an up-to-scale depth map through the process of Structure from Motion (SfM) [10]. Both approaches ultimately lead to the acquisition of discrete 3D points as separate voxels through a back-projection process based on the pinhole camera model, thereby realizing 3D reconstruction. Moreover, the approaches have been developed to the learning-based methods with significant advancements [7, 17]. Furthermore, endoscopic scenes have notable applications for depth estimation and 3D reconstruction [1, 12, 14]. However, a limitation of these evolved techniques is that discrete 3D points neglect the intricate topology of the scenes [26].

Different from generating discrete 3D points, neural radiation field (NeRF) leverages ray tracing to render RGB images from new views by means of continuous volumetric fields to achieve 3D reconstruction. This approach uses multiple posed images as input. Mildenhall et al. [16] firstly proposed NeRF to realize the view synthesis task with impressive achievements. Deng et al. [3] introduced the depth values to supervise the process of NeRF and improved results in fewer views. More recent works focus on enforcing the geometric constraints for NeRF. Xu et al. [25] built a semi-supervised framework for NeRF with geometry labels for depth values from multiple views. Choe et al. [2] introduced surface normal for surface regularization to improve the fidelity of NeRF. There are more approaches to modify NeRF by models such as vision transformer and diffusion model [13, 15]. NeRF is also applied to the surgical scenes. Gerats et al. [5] proposed a dynamic depth-supervised NeRF for the operating room. Wang et al. [22] realized the 3D reconstruction of the soft tissue by depth-supervised NeRF and offered endoscopic datasets. This method is further refined by Zha et al. [26], who introduced the surface normal to refine the 3D structures.

This work addresses a self-supervised monocular 3D scene reconstruction based on posed endoscopic videos with neural radiance fields. The previous methods [22, 26] were under a depth-supervised manner. And these approaches were reduced significantly without depth values as supervision during training time [22, 26]. For endoscopic datasets, collecting dense depth values is challenging due to the limited space and complex requirements of depth sensors [12].

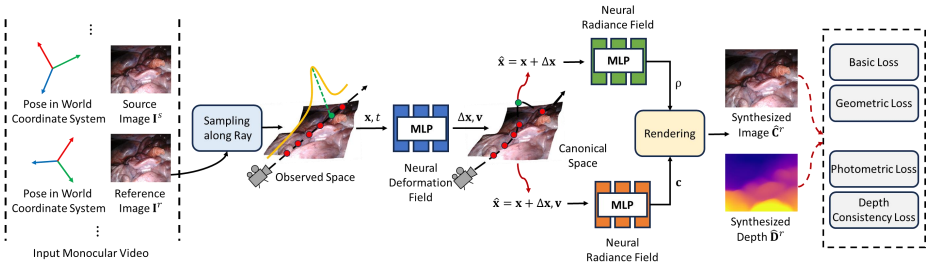

Fig. 1. Schematic flowchart of monocular 3D scene reconstruction of deformable tissues with neural radiance fields.

Also, the researchers neglect the temporal correlation between input images. We consider monocular 3D scene reconstruction by ray-tracing-based neural radiance fields and structure-from-motion-based photogrammetry. We utilize the structure from motion framework to complete the pixel-matching process based on the synthesized depth values and the endoscope's poses. Color is the only supervision signal in this process to complete a self-supervised learning strategy. In addition, we explicitly optimize the depth values obtained from the ray-tracing-based neural radiance field by introducing the geometric constraints between adjacent views. The locations of the corresponding depth values from adjacent views are also based on the back-projection process. To fully explore the temporal information, we propose a looped loss function for the whole optimization.

Our main contributions are summarized as follows. (i) We propose a self-supervised monocular 3D scene reconstruction with neural radiance fields on endoscopic videos to relax the dependence on depth in the previous methods. (ii) We explicitly introduce the geometric constraint on the predicted depth values from adjacent views. (iii) We propose a looped loss function for the optimization to explore the temporal information between the input images fully.

2 Method

2.1 Preliminaries

Problem Setting We aim to render RGB images from new views and reconstructions. Our method processes a series of data as $\{(\mathbf{I}^i, \mathbf{M}^i, \mathbf{P}^i, t^i)\}_{i=1}^{T}$. Here, T stands for the total number of frames. For each frame $\mathbf{I}^i \in \mathbb{R}^{H\times W\times 3}$ denotes the RGB image. Following the previous works [22], we also use the foreground mask $\mathbf{M}^i \in \mathbb{R}^{H\times W}$ to filter out irrelevant pixels. The timestamp $t^i = i/T$ represents the normalized time. $\mathbf{P}^i \in \mathbb{R}^{4\times 4}$ is the endoscope's pose corresponding to $\mathbf{I}^i$ from the view at time t^i. In addition, $\mathbf{K} \in \mathbb{R}^{4\times 4}$ is a matrix with the intrinsic parameters of the endoscope used in the self-supervised learning strategy.

Pipeline We implement a mask-guided sampling strategy [22] to respectively select valuable pixels from three input frames: reference image $\mathbf{I}^r$ and source im-

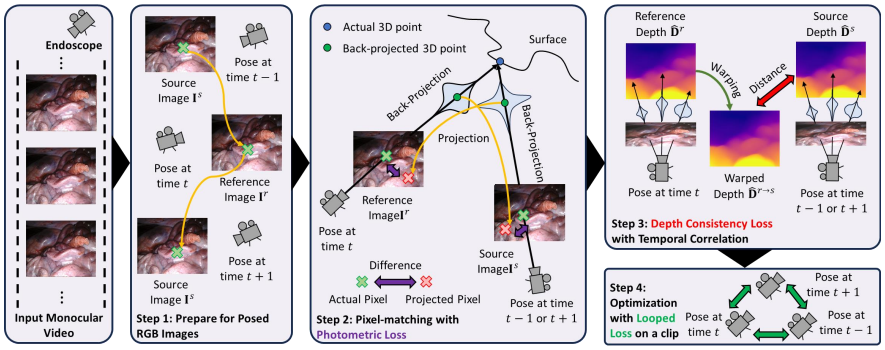

Fig. 2. Schematic process of self-supervised learning strategy to complete photometric loss and depth consistency loss with designed looped loss.

age $\mathbf{I}^s$. $\mathbf{I}^r$ is captured from the viewpoint at time t^r, and $\mathbf{I}^s$ is captured from the viewpoint at time t^s, including viewpoint at times t^{r-1} and t^{r+1}. $\mathbf{I}^r$ and $\mathbf{I}^s$ belong to $\mathbf{I}^i$. As shown in Fig. 1, we project 3D rays and sample points along the rays from these selected pixels. The multilayer perceptron (MLP) utilizes these sampled points to predict colors $\hat{\mathbf{C}}^r$ and $\hat{\mathbf{C}}^s$ with corresponding depths $\hat{\mathbf{D}}^r$ and $\hat{\mathbf{D}}^s$ by volume rendering method [21]. Then, we search the matched pixels between $\hat{\mathbf{C}}^r$ and $\mathbf{C}^s$ based on $\hat{\mathbf{D}}^r$ through structure from motion framework. This allows to generate the transformed color $\hat{\mathbf{C}}^{s\to r}$ to complete a self-supervised learning strategy. Inspired by [20], we tailor a looped loss function with photometric loss and depth consistency loss to train the networks. Fig. 2 shows the process of completing these loss functions under the self-supervised learning strategy.

2.2 Self-Supervised Learning with Depth Consistency

Structure from Motion Framework Inspired by the monocular depth estimation [27], we consider the self-supervised learning strategy as a view synthesis process based on pixel-matching process between adjacent images, as shown in Step 1 in Fig. 2. Following the pinhole camera model, we can back-project one pixel with 2D coordinates $\boldsymbol{p}$ to a 3D point in the current camera coordinate system by

$$\mathbf{Q}_{\boldsymbol{p}} = \mathbf{K}^{-1}\mathbf{D}_{\boldsymbol{p}}\boldsymbol{p}, \tag{1}$$

where $\mathbf{Q}$ is the point cloud included 3D points based on the back-projection process. $\mathbf{Q}_{\boldsymbol{p}}$ is a back-projected 3D point corresponding to the 2D position $\boldsymbol{p}$. Here, $\mathbf{Q}_{\boldsymbol{p}}$ is a vector containing three values. And $\boldsymbol{p}$ is represented with homogeneous coordinates in a vector containing three values. $\mathbf{K}$ is the intrinsic parameters of the endoscope, represented as a 3×3 matrix. $\mathbf{D}$ is a matrix that represents the depth map. $\mathbf{D}_{\boldsymbol{p}}$ is the depth value at 2D coordinates $\boldsymbol{p}$ in depth map $\mathbf{D}$, which is scalar value. Given the sampled pixels at 2D coordinates $\boldsymbol{p}^r$ in reference image $\mathbf{I}^r$, we obtain the 3D points from pixels at $\boldsymbol{p}^r$ through the back-projection process based on Eq. 1. The back-projected 3D points are in the

different coordinate systems from different views. Therefore, we utilize the poses of the endoscope to calculate the transformation matrix as the relative pose from the view at time t^r to the adjacent view at time t^s as $\mathbf{T}^{r\to s} = (\mathbf{P}^s)^{-1}\mathbf{P}^r$. Here, $\mathbf{P}^r$ and $\mathbf{P}^s$ are the poses of the endoscope. The corresponding pixels at 2D coordinates in source image $\mathbf{I}^s$ can be obtained by

$$\hat{\boldsymbol{p}}^s = \mathbf{K}\mathbf{T}^{r\to s}\hat{\mathbf{D}}^r_{\boldsymbol{p}^r}\mathbf{K}^{-1}\boldsymbol{p}^r, \tag{2}$$

where $\hat{\boldsymbol{p}}^s$ is the 2D coordinate of the pixel warped from the pixel at 2D coordinate $\boldsymbol{p}^r$ in reference image $\mathbf{I}^r$. $\hat{\mathbf{D}}^r_{\boldsymbol{p}^r}$ is the depth value at 2D coordinates $\boldsymbol{p}^r$ in the predicted depth map $\hat{\mathbf{D}}^r$, which is scalar value. We then generate the reference color based on the matched pixels $(\boldsymbol{p}^r, \hat{\boldsymbol{p}}^s)$ as $\mathbf{C}^{s\to r}_{\boldsymbol{p}^r} \leftarrow \mathbf{C}^s_{\hat{\boldsymbol{p}}^s}$. As shown in Step 2 in Fig. 2, we calculate the photometric correspondence as a supervised signal by

$$\mathrm{E}\left(\hat{\mathbf{C}}^{\mathrm{r}}, \mathbf{C}^{s\to r}, \boldsymbol{p}^r\right) = \alpha\frac{1-\mathrm{SSIM}\left(\hat{\mathbf{C}}^{\mathrm{r}}, \mathbf{C}^{s\to r}, \boldsymbol{p}^r\right)}{2} + (1-\alpha)\left|\hat{\mathbf{C}}^{\mathrm{r}}_{\boldsymbol{p}^r} - \mathbf{C}^{s\to r}_{\boldsymbol{p}^r}\right|, \tag{3}$$

where $\hat{\mathbf{C}}^{\mathrm{r}}$ is the synthesized color value based on the reference image's pose $\mathbf{P}^r$ and estimated color value along ray by MLP. α is set as 0.85 followed as [6] for structured similarity (SSIM) [23] and L1-norm operator $|\cdot|$. Furthermore, we utilize the minimum photometric error [6] in this self-supervised learning strategy by

$$\mathcal{L}_p = \frac{1}{N_m}\sum_{\boldsymbol{p}^r\in\mathcal{H}^r}\mathbf{M}_{\boldsymbol{p}^r}\left(\min_{\mathrm{s}}\mathrm{E}\left(\hat{\mathbf{C}}^{\mathrm{r}}_{\boldsymbol{p}^r}, \mathbf{C}^{s\to r}_{\boldsymbol{p}^r}\right)\right), \tag{4}$$

where r is the view of the reference image at time t^r, and s means the view of the source image at time t^s. Time t^s is time t^{r-1} or time t^{r+1}. $\mathbf{M}$ is the ray foreground mask. $\mathcal{H}^r$ is the set that includes pixels' coordinates in the $\hat{\mathbf{C}}^{\mathrm{r}}$. N_m is the number of valuable pixels selected by ray foreground mask $\mathbf{M}$.

Depth Consistency Loss NeRF-based approaches learn a scene's continuous volumetric density and color distribution using multi-view posed images as input to synthesize images from new views to complete 3D reconstruction. However, the previous method ignores the temporal correlation based on geometric constraints between input images [22, 26]. In the self-supervised learning strategy, we provide pixel-matching process based on the estimated depth and intrinsic parameters. As shown in Step 3 in Fig. 2, we provide consistency on synthesized depth from MLP through the pixel-matching process to provide multi-view geometric constraints by

$$\mathcal{L}_d = \frac{1}{N_m}\sum_{\boldsymbol{p}^r\in\mathcal{H}^r}\mathbf{M}_{\boldsymbol{p}^r}\left|\hat{\mathbf{D}}^{\mathrm{r}}_{\boldsymbol{p}^r} - \mathrm{Z}\left(\mathbf{K}\mathbf{T}^{s\to t}\hat{\mathbf{D}}^{\mathrm{s}}_{\hat{\boldsymbol{p}}^s}\mathbf{K}^{-1}\hat{\boldsymbol{p}}^s\right)\right|, \tag{5}$$

where $\hat{\mathbf{D}}^{\mathrm{r}}$ is the synthesized depths from reference image $\mathbf{I}^r$. $\hat{\mathbf{D}}^{\mathrm{r}}_{\boldsymbol{p}^r}$ is the depth value at the 2D coordinate $\boldsymbol{p}^r$ in $\hat{\mathbf{D}}^{\mathrm{r}}$. $\hat{\mathbf{D}}^{\mathrm{s}}_{\hat{\boldsymbol{p}}^s}$ is the depth value at the 2D coordinate $\hat{\boldsymbol{p}}^s$ in $\hat{\mathbf{D}}^{\mathrm{s}}$. $\boldsymbol{p}^r$ and $\hat{\boldsymbol{p}}^s$ are matched pixels based on the Eq. 2. $\mathbf{T}^{s\to r}$ is the

transformation matrix from the view at time s to the view at time s defined by $\mathbf{T}^{s\to r} = (\mathbf{P}^r)^{-1}\,\mathbf{P}^s$. Operator Z extracts the value in the third channel. Here, we obtain the value in the z-axis from the 3D points back-projected at $\hat{\boldsymbol{p}}^s$.

2.3 Optimization

Rendering Given an endoscope's pose $\mathbf{P}^i = \left[\mathbf{R}^i|\mathbf{t}^i\right]$, each pixel's 2D coordinate $\boldsymbol{p} \in \mathbb{R}^2$ determines a ray in the world coordinate system, whose origin is the endoscope center of projection $\mathbf{o}^i = \mathbf{t}^i$ and whose direction is defined as $\mathbf{v}^i_{\boldsymbol{p}} = \mathbf{R}^i\mathbf{K}^{-1}\boldsymbol{p}$. We sample 3D point along the viewing ray $\mathbf{y}$ associated with $\boldsymbol{p}$ at depth candidate z^n as $\mathbf{y}^{i,n}_{\boldsymbol{p}} = \mathbf{o}^i + z^n\mathbf{v}^i_{\boldsymbol{p}}$. Here, n is the index of the depth candidate. The color $\hat{\mathbf{C}}$ and depth $\hat{\mathbf{D}}$ of the ray can be approximated by

$$\hat{\mathbf{C}}_{\mathbf{y}} = \sum\nolimits_{i=1}^{N^i} T^i\alpha^i\mathbf{c}^i, \quad \hat{\mathbf{D}}_{\mathbf{y}} = \sum\nolimits_{i=1}^{N^i} T^i\alpha^i z^i, \tag{6}$$

where $T^i = \prod_{j=1}^{i-1}(1-\alpha^j)$, $\alpha^i = \max((\phi(\rho^i) - \phi(\rho^{i+1}))/\phi(\rho^i), 0)$ and $\phi(\rho) = (1+e^{-\rho/s})^{-1}$. $\mathbf{c}$ is the predicted color as the output of the MLP.

Loss Function We use training objectives as basic and self-supervised loss functions. The basic loss function minimizes the difference between the actual values $\mathbf{C}^r$ and rendered values $\hat{\mathbf{C}}^r$ as $\mathcal{L}_b = \frac{1}{N_m}\sum_{\boldsymbol{p}^r\in\mathcal{H}^r} \|\mathbf{M}_{\boldsymbol{p}^r}\left(\hat{\mathbf{C}}^r_{\boldsymbol{p}^r} - \mathbf{C}^r_{\boldsymbol{p}^r}\right)\|_1$. In addition, we use the geometric loss $\mathcal{L}_g$ as one of the basic terms in the previous method [26]. We use the predicted depth value instead of the actual depth value as ground truth for the geometric loss $\mathcal{L}_g$. The self-supervised loss function completes the self-supervised learning strategy by structure from motion framework with depth consistency to provide temporal correspondence.

To enhance the utilization of temporal correlation, we treated each input image as a reference image from a clip of the monocular video to establish a looped prediction learning framework, as shown in Step 4 in Fig. 2. The overall loss in the final computation is derived as an average of the errors from each individual combination as $\mathcal{L}_f = \frac{1}{N_k}\sum_{k=1}^{N_k}\mathcal{L}^k_b + \lambda\mathcal{L}^k_g + \gamma\mathcal{L}^k_p + \delta\mathcal{L}^k_d$. k is the index of reference view. N_k is the number of input images, and N_k equals 3.

3 Experiments

3.1 Experiment Settings

Datasets and Evaluation Our research involves experiments on two publicly available endoscope datasets: ENDONERF [22] and SCARED [1]. The ENDONERF dataset provides two examples of in-vivo prostatectomy data, complete with manually labeled foreground masks. The SCARED dataset consists of 35 endoscopic videos taken from 9 various scenes. To prepare the data for our experiments, we followed a process established in the previous method [26]. Also, we

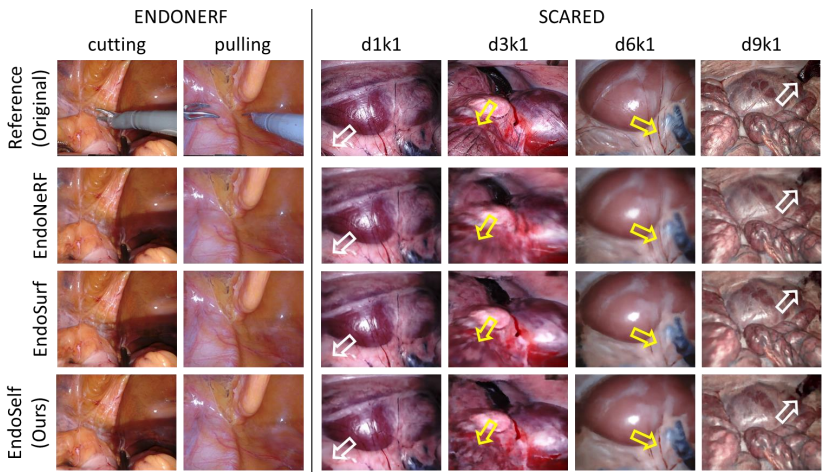

Fig. 3. Comparison of rendered RGB images. The first row shows the original image as color ground truth. Our method renders RGB images with more details (yellow arrows). Previous methods lost the tissue structures in the corners (white arrows).

adopted all the scenes in ENDONERF and SCARED to conduct the comparison experiments. This involves adjusting the scenes within each dataset to fit within a unit sphere, ensuring consistency across all data. We then divided the frame data into 7:1 training and test sets. ENDONERF included 219 and 28 frames for training and testing. SCARED included 2,434 and 307 frames for training and testing. Due to the static endoscopic position, we randomly added perturbation into the endoscope poses in ENDONERF. To evaluate the quality of rendered RGB images based on this data, we used three standard metrics: PSNR, SSIM, and LPIPS. Since this work was under the self-supervised learning strategy, the experiments focused on the evaluation of 2D synthesized images.

Implementation Details We trained individual neural network models for each scene. These networks have 8 layers with 256 channels each, including a skip connection at the 4th layer. We utilized PyTorch [19] with the Adam optimizer [11] with a learning rate of 0.0005, starting with a 5,000 iteration warm-up before decaying at a rate of 0.05. Each training batch comprises 1,024 rays and 64 points per ray, with an initial standard deviation of 0.3. The loss function weights are $\lambda = 0.25$, $\gamma = 0.2$, and $\delta = 0.001$. The training runs for 100,000 iterations, taking 17 hours on an NVIDIA A100 GPU.

3.2 Comparison Evaluation

We re-trained all the models of approaches for each scene. We compared the proposed method with existing NeRF-based methods [22, 26]. Table 1 showed the quantitative results for the rendered RGB images based on three widely used

metrics. As listed in Table 1, EndoSelf had better results compared to EndoNeRF [22] and EndoSurf [26] on each scene of ENDONERF [22] and SCARED [1]. And we calculate the average results for each dataset as shown in Table 1. Comparing to EndoSurf, EndoSelf produces better results than by ↑ 0.598 PSNR, ↑ 0.001 SSIM, and ↑ 0.009 LPIPS on ENDONERF [22]. And EndoSelf had better results by ↑ 0.434 PSNR, ↑ 0.007 SSIM, and ↑ 0.008 LPIPS on SCARED [1]. Note that EndoNeRF and EndoSurf are both depth-supervised approaches and EndoSelf are trained under a self-supervised manner without depth value as a supervised signal. Each method performs with high scores on ENDONERF compared to the performance of results on SCARED. For ENDONERF, there is very slight variation within each scene, making reaching a better view synthesis easier. Hence, we adopted all 9 scenes in SCARED to provide more comparisons. Fig. 3 also shows each method behaves similarly on ENDONERF for qualitative results. However, SCARED includes Out-of-View Movement between adjacent frames. Our method leverages temporal correlation based on the back-projection process to provide more details (yellow arrow) and missing objects (white arrow).

3.3 Ablation Study

We performed the ablation study on other proposed components' contribution to the rendering in Table 2. The components involve Consistency on Photometric Error (CPE); Consistency on Depth Value (CDV); and Looped Lose Function (LLF). The ablation study reveals the performance became slightly worse when converting the depth-supervised learning strategy to a self-supervised learning strategy on ENDONERF [22] and SCARED [1] (IDs 1, 2, 3 and 4). However, the whole proposed method outperformed better than the previous depth-supervised approach (IDs 1 and 5). Each component contributes to the proposed method (IDs 2, 3 and 4). Furthermore, the proposed method improves significantly when combining each component (IDs 2, 3, 4 and 5).

Table 1. Quantitative comparison for renderer RGB images with three metrics on ENDONERF and SCARED. The best performance is bold.

Methods		EndoNeRF			EndoSurf			EndoSelf (Ours)		
Metrics		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
ENDONERF	cutting	34.242	0.932	0.151	34.896	0.952	0.107	**35.439**	**0.953**	**0.103**
	pulling	34.188	0.938	0.160	34.917	0.955	0.121	**35.651**	**0.958**	**0.112**
	Average	34.227	0.934	0.153	34.902	0.953	0.111	**35.500**	**0.954**	**0.106**
SCARED	d1k1	24.669	0.768	0.351	24.666	0.777	0.340	**25.031**	**0.781**	**0.336**
	d2k1	25.296	0.829	0.262	26.125	0.836	0.263	**26.421**	**0.839**	**0.246**
	d3k1	19.858	0.619	0.466	21.593	0.682	0.417	**22.084**	**0.689**	**0.412**
	d4k1	22.354	0.796	0.423	22.774	0.835	0.386	**23.096**	**0.851**	**0.363**
	d5k1	22.429	0.806	0.385	23.089	0.865	**0.296**	**23.514**	**0.870**	0.311
	d6k1	24.745	0.848	0.472	25.171	0.882	0.435	**25.800**	**0.885**	**0.433**
	d7k1	23.221	0.840	0.299	24.686	0.886	**0.247**	**25.332**	**0.890**	0.248
	d8k1	24.611	0.800	0.489	25.511	0.834	0.451	**25.735**	**0.844**	**0.425**
	d9k1	22.080	0.633	0.518	22.382	0.658	**0.472**	**22.771**	**0.662**	**0.472**
	Average	23.008	0.765	0.424	23.792	0.803	0.382	**24.226**	**0.810**	**0.374**

Table 2. Evaluation for variants of the proposed model on ENDONERF and SCARED. DS: Depth Supervision Signal ; CPE: Consistency on Photometric Error; CDV: Consistency on Depth Value; LLF: Looped Lose Function. The best performance is bold.

ID	DSS	CPE	CDV	LLF	ENDONERF			SCARED		
					PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
1	✓				34.902	0.953	0.111	23.792	0.803	0.382
2		✓			34.600	0.946	0.124	23.960	0.807	0.382
3		✓	✓		34.681	0.947	0.124	24.026	0.808	0.380
4		✓		✓	34.652	0.947	0.123	24.024	0.808	0.379
5		✓	✓	✓	**35.500**	**0.954**	**0.106**	**24.226**	**0.810**	**0.374**

4 Conclusions

This study proposes EndoSelf, a self-supervised approach based on neural fields to reconstruct deforming surgical scenes from monocular endoscopic videos. Unlike previous methods that relied on prior depth values as supervised signals, Endoself overcomes the limitation of the prior depth value as one of the supervised signals based on the back-projection geometry of photogrammetry. In addition, we propose depth consistency and looped loss function to leverage the temporal correlation between adjacent images fully. Experimental results on two public datasets showed that our method outperformed previous methods. In the future, we will utilize fewer posed images to complete the view synthesis and reconstruction because collecting accurate poses is hard for endoscopic scenes.

Acknowledgments. This work was funded by grants from the JSPS KAKENHI (24H00720, 24K03262), the JST CREST (JPMJCR20D5), the JST [Moonshot R&D] (JPMJMS2033, JPMJMS2214), and the JSPS Bilateral Joint Research Project.

Disclosure of Interests. The authors report there are no competing interests to declare.

References

1. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. Choe, J., Choy, C., Park, J., Kweon, I.S., Anandkumar, A.: Spacetime surface regularization for neural dynamic scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17871–17881 (2023)
3. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
4. Fuchs, H., Livingston, M.A., Raskar, R., Colucci, D., Keller, K., State, A., Crawford, J.R., Rademacher, P., Drake, S.H., Meyer, A.A.: Augmented reality visualization for laparoscopic surgery. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI'98: First International Conference Cambridge, MA, USA, October 11–13, 1998 Proceedings 1. pp. 934–943. Springer (1998)

5. Gerats, B.G., Wolterink, J.M., Broeders, I.A.: Dynamic depth-supervised NeRF for multi-view RGB-D operating room videos. In: International Workshop on PRedictive Intelligence In MEdicine. pp. 218–230. Springer (2023)
6. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
7. Hamid, M.S., Abd Manap, N., Hamzah, R.A., Kadmin, A.F.: Stereo matching algorithm based on deep learning: A survey. Journal of King Saud University-Computer and Information Sciences **34**(5), 1663–1673 (2022)
8. Hirschmuller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 807–814. IEEE (2005)
9. Hu, M., Penney, G., Edwards, P., Figl, M., Hawkes, D.J.: 3D reconstruction of internal organ surfaces for minimal invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007. pp. 68–77. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
10. Iglhaut, J., Cabo, C., Puliti, S., Piermattei, L., O'Connor, J., Rosette, J.: Structure from motion photogrammetry in forestry: A review. Current Forestry Reports **5**, 155–168 (2019)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Li, W., Hayashi, Y., Oda, M., Kitasaka, T., Misawa, K., Mori, K.: Multi-view guidance for self-supervised monocular depth estimation on laparoscopic images via spatio-temporal correspondence. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 429–439. Springer Nature Switzerland, Cham (2023)
13. Lin, K.E., Lin, Y.C., Lai, W.S., Lin, T.Y., Shih, Y.C., Ramamoorthi, R.: Vision transformer for NeRF-based view synthesis from a single input image. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 806–815 (2023)
14. Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., et al.: Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. Medical image analysis **17**(8), 974–996 (2013)
15. Melas-Kyriazi, L., Laina, I., Rupprecht, C., Vedaldi, A.: RealFusion: 360deg reconstruction of any object from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8446–8455 (2023)
16. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
17. Ming, Y., Meng, X., Fan, C., Yu, H.: Deep learning for monocular depth estimation: A review. Neurocomputing **438**, 14–33 (2021)
18. Nisky, I., Huang, F., Milstein, A., Pugh, C.M., Mussa-Ivaldi, F.A., Karniel, A.: Perception of stiffness in laparoscopy–the fulcrum effect. Studies in health technology and informatics **173**, 313 (2012)
19. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS 2017 Workshop on Autodiff (2017)
20. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4190–4200 (2023)

21. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
22. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 431–441. Springer Nature Switzerland, Cham (2022)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
24. Xia, J., Samman, N., Chua, C.K., Yeung, R.W., Wang, D., Shen, S.G., Ip, H.H., Tideman, H.: PC-based virtual reality surgical simulation for orthognathic surgery. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000: Third International Conference, Pittsburgh, PA, USA, October 11-14, 2000. Proceedings 3. pp. 1019–1028. Springer (2000)
25. Xu, D., Jiang, Y., Wang, P., Fan, Z., Shi, H., Wang, Z.: Sinnerf: Training neural radiance fields on complex scenes from a single image. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
26. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: EndoSurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 13–23. Springer Nature Switzerland, Cham (2023)
27. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017)