ColonNeRF: Neural Radiance Fields for High-Fidelity Long-Sequence Colonoscopy Reconstruction

Yufei Shi[1*], Beijia Lu[1*], Jia-Wei Liu[12†], Ming Li[2], Mike Zheng Shou[1✉],
[1]Show Lab, [2]Institute of Data Science, National University of Singapore

Abstract

Colonoscopy reconstruction is pivotal for diagnosing colorectal cancer. However, accurate long-sequence colonoscopy reconstruction faces three major challenges: (1) dissimilarity among segments of the colon due to its meandering and convoluted shape; (2) co-existence of simple and intricately folded geometry structures; (3) sparse viewpoints due to constrained camera trajectories. To tackle these challenges, we introduce a new reconstruction framework based on neural radiance field (NeRF), named ColonNeRF, which leverages neural rendering for novel view synthesis of long-sequence colonoscopy. Specifically, to reconstruct the entire colon in a piecewise manner, our ColonNeRF introduces a region division and integration module, effectively reducing shape dissimilarity and ensuring geometric consistency in each segment. To learn both the simple and complex geometry in a unified framework, our ColonNeRF incorporates a multi-level fusion module that progressively models the colon regions from easy to hard. Additionally, to overcome the challenges from sparse views, we devise a DensiNet module for densifying camera poses under the guidance of semantic consistency. We conduct extensive experiments on both synthetic and real-world datasets to evaluate our ColonNeRF. Quantitatively, our ColonNeRF outperforms existing methods on two benchmarks over four evaluation metrics. Notably, our LPIPS-ALEX scores exhibit a substantial increase of about 67%-85% on the SimCol-to-3D dataset. Qualitatively, our reconstruction visualizations show much clearer textures and more accurate geometric details. These sufficiently demonstrate our superior performance over the state-of-the-art methods. Our project page is available at https://showlab.github.io/ColonNeRF/.

1. Introduction

Colorectal cancer (CRC) is a leading cause of death, ranking fourth only after lung, breast, and prostate cancer [2]. Despite its prevalence, the 5-year survival rate can rise to 90% for those who undergo early screening [16]. Therefore, identifying colorectal cancer in the early stage is essential [33, 1, 25]. Colonoscopy [10] has become one of the most crucial examinations for the early diagnosis of CRC due to its convenient operations and effectiveness.

However, the preciseness of colonoscopy scans is still limited by the intricate geometry of the colon. It is reported that even experienced physicians are likely to overlook about 22-28% of polyps since they only rely on 2D scans without any 3D details [19]. Therefore, high-fidelity colonoscopy reconstruction is critical for CRC diagnosis. The reconstruction is also a prerequisite for various downstream clinical applications, *e.g.* preoperative review and surgical planning [39]. Additionally, it is an important tool for medical education and offers hands-on training and skill development.

Traditional methods such as SLAM [7] have been introduced into colonoscopy reconstruction by matching two-dimensional (2D) image pixels and their corresponding 3D spatial points in endoscopic scenes. Specifically, Ma et al. [27] combine a standard SLAM system with a depth and pose prediction network and achieve a robust tracking system. Meanwhile, Wang et al. [38] utilize the characteristics of surface normal vectors to develop a two-step neural framework as initialization for a SLAM-based pipeline to improve the reconstruction quality. However, despite the capability of SLAM in constructing environmental maps and tracking agent locations [12], it falls short when tasked with novel view synthesis, which necessitates a comprehensive understanding of the 3D structure of the scene. As a result, SLAM fails to produce a comprehensive 3D reconstruction, limiting its practical use in real-world scenarios.

To address the novel view synthesis issue in 3D reconstruction, NeRF [27] proposes the neural implicit field for continuous scene representations and achieves great success in producing high-quality novel view images for complicated scenarios. Inspired by these, EndoNeRF [39], the first work leveraging neural rendering (NeRF), exhibits great performance in 3D reconstruction and deformation track-

* Equal Contribution † Project Lead ✉ Corresponding Author

ing of surgical scenes. Unlike EndoNeRF, which focuses on limited scene reconstruction, our principal objective is the precise reconstruction of entire long-sequence colonoscopies. Till now, several key challenges of colonoscopy reconstruction remain unsolved.

Firstly, the inherent meandering and convoluted shape of the colon results in dissimilarity across its different segments. This variability poses significant obstacles for achieving high-quality reconstruction of long-sequence colonoscopy when directly utilizing NeRF. Secondly, the co-existence of simple and intricately folded geometry structure makes it difficult for the model to adequately learn every detail in the segment, resulting in losses of crucial details in the imaging data and posing significant challenges for accurate colonoscopy reconstruction. Lastly, due to the constrained camera trajectory during colonoscopy capturing, the colonoscopy data is featured with sparse viewpoints [39], hindering the performance of previous methods [27].

To resolve the challenges mentioned above, we propose a new model for 3D colonoscopy reconstruction dubbed ColonNeRF. The ColonNeRF comprises three purpose-built modules. Each module in ColonNeRF is meticulously designed to address a specific reconstruction challenge and they are elegantly combined, ensuring comprehensive and accurate colonoscopy reconstruction. To overcome the first challenge of dissimilarity among colon regions due to its meandering and convoluted shape, we design a region division and integration module to ensure geometric consistency in each unit. Specifically, the division module is utilized to divide the colon into multiple segments in a soft manner, which is based on the curvature of the colon with joint regions learned by both adjacent segments. Then the integration module is responsible for fusing all divided segments under the double-filtering strategy. To learn both simple and complex geometry in a unified framework, we use a multi-level fusion module to progressively model the colon structures, enhancing textural and geometric details in a coarse-to-fine way. To deal with the challenge from sparse viewpoints, we design a DensiNet module in each stage to encourage our model to learn colon features from three angles: original pose, spinning around pose, and helix rotating pose. Specifically, we employ the DINO-ViT-based semantic consistency regularization to supervise the reconstruction from densified camera poses.

1.1. Contributions

In our work, we make the following main contributions:

- We design ColonNeRF, a new 3D framework leveraging neural rendering for high-quality long-sequence colonoscopy reconstruction.
- We design a series of purpose-built reconstruction mechanisms consisting of the region division and integration module, multi-level fusion module, and DensiNet module. These proposals are demonstrated critical for successfully achieving our state-of-the-art synthesis results.
- Experiments on synthetic and real-world datasets demonstrate that our method achieves high-quality novel view synthesis of long-sequence colonoscopy, which outperforms the baseline methods. Notably, it achieves significant improvements of about 21%-29% and 67%-85% in terms of the LPIPS-VGG and LPIPS-ALEX metrics respectively on the synthetic dataset.

1.2. Paper Organization

The rest of the paper is organized as follows. In Section II, we provide a review of the most related work, focusing on recent developments in 3D reconstruction and colonoscopy reconstruction. Section III introduces the preliminary knowledge of Neural Radiance Fields (NeRF), which forms the foundation of our study. Section IV is dedicated to a comprehensive description of our proposed approach, detailing the methodologies and techniques employed. Section V involves a detailed quantitative and qualitative evaluation of our model, conducted using two distinct datasets to ensure robust assessment. Finally, Section VI concludes the paper with a comprehensive summary of our research work.

2. Related Works

2.1. Advances in 3D Reconstruction

Many studies in the field of 3D scene reconstruction and novel view synthesis (NVS) have traditionally relied on methods such as Lumigraph [17], light field functions [20], meshes [14], voxels [3], point clouds [13], and Multi-Plane Images (MPI) [9]. While these techniques have significantly advanced the field, each has inherent limitations that affect their application in complex scenarios. For example, Lumigraph and light field functions exhibit poor continuity between different viewpoints and struggle with complex lighting conditions and shadow effects, which limits their practical application in the long-sequence colon datasets. Meshes require a complex topological structure unsuitable for modeling intricate colon scenes. Due to their discrete sampling (synthesizing higher resolution images needs a finer sampling of 3D space), the voxels and point clouds are limited by poor time and space complexity. Multi-Plane Images (MPI) encounter difficulties in modeling complex geometrical shapes and occlusions [30].

Neural Radiance Fields (NeRF) [28] achieve impressive results in novel view synthesis by learning implicit neural scene representations. Since its emergence, numerous advancements have been made to break through

the limited performance of traditional 3D reconstruction by leveraging differentiable rendering and neural networks [37, 8, 15, 23, 21] for high-fidelity novel view synthesis of static and dynamic scenes. For instance, Xu et al. [41] utilize pseudo-labels on unseen viewpoints to guide the training process and improve model performance. Barron et al. [4] propose casting a conical frustum instead of a single ray to solve the anti-aliasing problem. Extensions of NeRF now address complex and large environments, as demonstrated by Yuanbo et al. [34], who developed a variant for large-scale scene rendering, and BungeeNeRF [40], which offers multiscale rendering. The versatility of NeRF has also expanded into generation and editing applications [29, 43, 22], underscoring the remarkable progress of NeRF in 3D reconstruction and novel view synthesis.

2.2. Colonoscopy Reconstruction

Previous works have explored 3D colonoscopy reconstruction based on the above 3D representations. Ma et al. [26, 27] develop a SLAM-based system with a post-averaging step to correct camera pose errors, showcasing advancements in camera tracking. In addition, Rau et al. [31] leverage SFM pseudo-labels and RNN models for 6D camera pose prediction, integrating deep learning into reconstruction. Wang et al. [38] utilize the relationship between illumination and surface normals to refine the normal and depth predictions recursively. Liu et al. [24] propose a SLAM system with appearance and geometry prior to reconstruct the 3D geometry of the observed region. While SLAM excels in generating environmental maps and tracking the spatial positioning of agents, its performance is compromised during novel view synthesis due to the necessity for detailed modeling of the scene 3D structure. This deficiency impedes the model to deliver an exhaustive 3D reconstruction, thus constraining its utility in practical real-world settings. The introduction of NeRF [28] marked a turning point, leading to methods like EndoNeRF [39], which utilize neural rendering for surgical scene reconstructions. However, EndoNeRF [39] focuses on limited scene reconstruction and is unsuitable for long-sequence colon reconstruction.

2.3. Preliminaries

Neural Radiance Fields (NeRF) [28] synthesize novel views of a scene by mapping 5D coordinates, comprising 3D position $\mathbf{x}$ and 2D viewing direction $\mathbf{d}$ to RGB color $\mathbf{c}$ and volumetric density σ. Each pixel in an image corresponds to a ray $\mathbf{r}(\tau) = \mathbf{o} + \tau\mathbf{d}$, where $\mathbf{o}$ is the camera origin, and $\mathbf{d}$ is the ray direction, τ is the distance between the origin point and sample point. The predicted color $\mathbf{C}(\mathbf{r})$ of the pixel can be represented as:

$$\mathbf{C}(\mathbf{r}) = \int_{\tau_{\text{near}}}^{\tau_{\text{far}}} T(\tau)\sigma(\mathbf{r}(\tau))\mathbf{c}(\mathbf{r}(\tau), \mathbf{d})d\tau \ , \qquad (1)$$

$$\text{where } T(\tau) = \exp\left(-\int_{\tau_{\text{near}}}^{\tau} \sigma(\mathbf{r}(s))ds\right). \qquad (2)$$

To facilitate the NeRF Multilayer Perceptrons (MLPs) in capturing more high-frequency details [35], the inputs $\mathbf{x}$ and $\mathbf{d}$ are each preprocessed through a sinusoidal positional encoding γ:

$$\gamma(z) = \left[\sin(z), \cos(z), \ldots, \sin\left(2^{L-1}z\right), \cos\left(2^{L-1}z\right)\right]^{\mathrm{T}} \qquad (3)$$

where L is the number of levels of positional encoding.

The NeRF [28] model optimizes the radiance field by minimizing the mean squared error between the synthetically rendered color and the ground truth color, as given by:

$$\mathcal{L}_{\text{pixel}} = \textstyle\sum_{\mathbf{r}\in R_i} \|(\mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r}))\|^2 \qquad (4)$$

where R_i is the set of input rays during training, $\hat{\mathbf{C}}(\mathbf{r})$ and $\mathbf{C}(\mathbf{r})$ is the ground truth and predicted RGB colors for ray $\mathbf{r}$.

3. Methodology

3.1. Framework Architecture

As shown in Fig. 1, given long-sequence colonoscopy data, we first split the data by region division module (Sec. B) to ensure geometric consistency within each segment. To learn both the simple and complex geometry in a unified framework, we use the multi-level fusion module (Sec. C) to progressively learn the colon geometry structure, improving the texture and geometry details in a coarse-to-fine way. Drawing inspiration from the BungeeNeRF [40], the model adopts the residual connection to enable gradients obtained from the latter MLPs to flow back to earlier MLPs smoothly. Subsequently, DensiNet module (Sec. D) tackles sparse data by densifying camera poses, incorporating original, spinning around, and helix rotating poses to augment data.

As shown in Fig. 2, during rendering, we run the region integration module (Sec. E) to filter out blocks that contribute minimally to the final output and integrate blocks containing pertinent information to ensure a seamless transition between blocks. Finally, we summarize the training objectives (Sec. F).

3.2. Region Division Module

To address the inherent dissimilarities in different colon segments that are characterized by varying diameters and curvatures, we develop a region division module for the colon's meandering and convoluted structure. This module aims to reconstruct the entire colon piecewise, reducing shape dissimilarity and ensuring geometric consistency in each segment. Specifically, it segments the colon into blocks at bends or locations with significant angle changes.

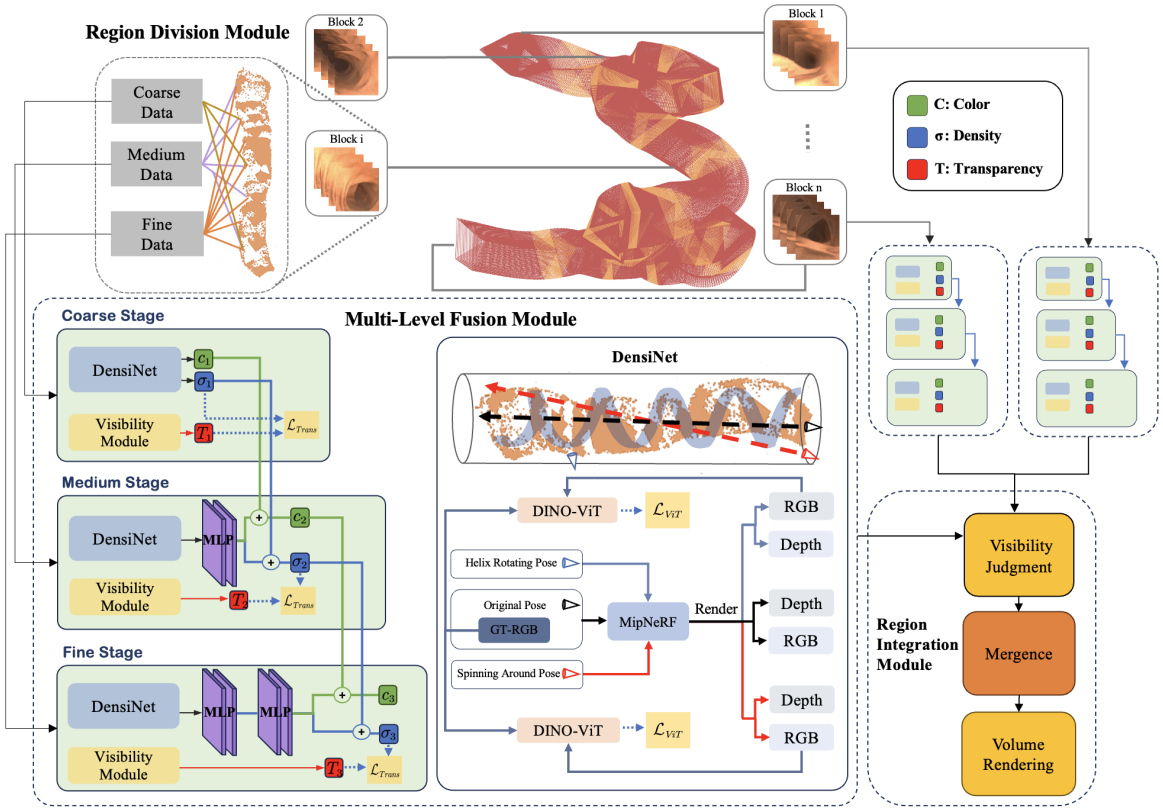

Figure 1. Overview of ColonNeRF. The architecture comprises a region division module, depicted in the upper section, where orange areas illustrate transition zones between adjacent regions. Each region includes a core area (red) and an adjacent transition zone, processed through various sparsity levels to produce coarse, medium, and fine data. These data feed into a multi-level fusion module, with each stage containing an DensiNet module for data augmentation. Within DensiNet module, we input the helix rotating pose, original pose, and spinning around pose into the MipNeRF [4] to optimize intestinal geometry learning. A DINO-ViT module is included for supervised training. Following processing through this module, final color, density, and transparency are determined, and the region integration module executes information filtering, fusion, and rendering across all blocks.

This approach not only promotes shape similarity within each segment but also surpasses traditional methods that process the colon as a single unit, enhancing the overall quality and accuracy of the reconstruction.

Applying this region division module to our datasets, we adapt its segmentation strategy to suit each dataset's specific geometry characteristics. In the synthetic dataset, the module divides the colon into 31 distinct blocks, with each block containing approximately $40 \sim 50$ images. For the real-world dataset, we divide it into four blocks, each comprising $17 \sim 19$ images. We ensure a 30% overlap between adjacent blocks to maintain seamless transitions, a critical aspect for accurate reconstruction. This overlapping strategy is illustrated in Fig. 1, where each block is represented with a central red region surrounded by two orange regions, indicating the areas of overlap. This methodological approach, detailed further in our ablation study, ensures a more accurate reconstruction of the colon complex geometry.

3.3. Multi-Level Fusion Module

Given the geometry of the colon, with its blend of simple surfaces, intricate folds, and numerous blood vessels and protrusions, the complexity of model reconstruction is significantly heightened. Relying solely on single-scale inputs, which focus on a specific scale, is insufficient for capturing the full spectrum of features. To surmount this challenge, we design a multi-level fusion module that progressively models the colon structures, enhancing textural and geometric details in a coarse-to-fine manner.

Specifically, the multi-level fusion module initiates with inputs of low sparsity RGB, depth, and pose data. It progressively incorporates denser data, enabling a smooth transition from coarse to fine details, thus enhancing the effectiveness of the feature extraction process. As the model advances to the next stage, we integrate additional Multilayer Perceptron (MLP) modules, as demonstrated in Fig. 1. The level of data sparsity at each ith stage of the input model

is calculated using the formula $\frac{2^n}{T*2^i}$, where i denotes the stage number, ranging from one to n, and T represents the total duration of the detection.

Each stage of the module includes two sub-modules: DensiNet and the visibility module. DensiNet generates RGB and density σ values for each spatial position, while the visibility module, comprising a four-layer MLP network and a linear output layer, calculates the transparency T_i for each spatial ray. The visibility module supervises transparency with the density σ output from DensiNet, following the formula to calculate the transparency loss:

$$\mathcal{L}_{\text{trans}} = \parallel T_{\text{i}} - \sigma_{\text{i}} \parallel \tag{5}$$

As the model progresses, it inherits the parameters of DensiNet and the visibility module from the previous stage, adding two residual connections to link the color and density outputs from the previous stage to the next. The final output combines the newly calculated RGB c_2 and density σ_2 values with outputs from each stage, resulting in a comprehensive final image.

$$\sigma_{\text{output}} = \sigma_{\text{L}}(\textstyle\sum_{\sigma_1}^{\sigma_{\text{n}}}) \qquad C_{\text{output}} = \zeta(\textstyle\sum_{\text{C}_1}^{C_{\text{n}}}) \tag{6}$$

The activation functions applied to the final values of σ and c include the Sigmoid function σ_{L} for density and a Softplus function ζ for color. This architectural design is proficient in integrating features at varying sparsity levels, thereby facilitating a detailed and nuanced image restoration.

3.4. DensiNet Module

Due to the constrained camera movement trace during the colonoscopic sampling, the sparse 3D points acquired can significantly deteriorate the quality of the reconstruction. To deal with this challenge, we design the DensiNet module, which leverages MipNeRF [4] as its backbone. Although MipNeRF [4] is superior to the original NeRF [28] in handling ambiguity, it still struggles with sparse data sampling. Our DensiNet module enhances the ability of the model to capture the colon features from sparse camera poses.

In the DensiNet module, our approach begins with patch sampling from RGB and depth images under the original view. Specifically, we extract 56×56 patches using a stride of 7 and calculate patch loss by comparing the difference between these extracted patches and their counterparts in the post-rendering images using the formula below.

$$\mathcal{L}_{\text{patch}} = \mathcal{L}_{\text{p}}(\boldsymbol{R}_1, f(\boldsymbol{R}_1)) + \mathcal{L}_{\text{p}}(\boldsymbol{D}_1, f(\boldsymbol{D}_1)) \tag{7}$$

where $\mathcal{L}_{\text{p}}$ represents the patch loss. $\boldsymbol{R}_1$ and $\boldsymbol{D}_1$ represent the sampled points of RGB and depth image obtained through the patch sampling technique. The function f corresponds to the processing carried out by the MipNeRF [4] network. $f(\boldsymbol{R}_1)$ and $f(\boldsymbol{D}_1)$ refer to the RGB and depth output results from the MipNeRF [4].

Secondly, to further improve the capability to learn structure from the original viewpoint, we randomly select 3,136 points from the RGB and depth images. We compute the Mean Squared Error (MSE) loss between these points and their corresponding post-rendered points from RGB and depth rendering results, as the subsequent formula defines.

$$\mathcal{L}_{\text{rand}} = \mathcal{L}_{\text{m}}(\boldsymbol{R}_2, f(\boldsymbol{R}_2)) + \mathcal{L}_{\text{m}}(\boldsymbol{D}_2, f(\boldsymbol{D}_2)) \tag{8}$$

where $\mathcal{L}_{\text{m}}$ represents the Mean Squared Error (MSE) loss. The variables $\boldsymbol{R}_2$ and $\boldsymbol{D}_2$ correspond to the points sampled from the RGB and depth images using a random selection strategy. And we could get the final original pose loss.

$$\mathcal{L}_{\text{ori}} = \mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{rand}} \tag{9}$$

To counter the sparsity of data, we integrate supervision from two novel poses - the spinning around pose and the helix rotating pose. These poses, designed to explore the surrounding region of the original pose and the colon wall's geometric structure, respectively, enhance the model's capacity for semantic consistency. We elucidate the specifics of these poses in subsequent sections.

Spinning Around Pose. To enhance the reconstruction of geometric structures around the original pose, we employ a rotation transformation to obtain spinning around pose from the original pose. For any given pixel $\boldsymbol{P}(x_{\text{i}}, y_{\text{i}})$ on the original view, its corresponding position on the destination pose $\boldsymbol{P}_{\text{des}}$ can be represented as:

$$\boldsymbol{P}_{\text{des}} = \begin{bmatrix} \boldsymbol{R}_{\text{des}} & \boldsymbol{t}_{\text{des}} \\ \boldsymbol{0} & \boldsymbol{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{\text{ori}} & \boldsymbol{t}_{\text{ori}} \\ \boldsymbol{0} & \boldsymbol{1} \end{bmatrix}^{-1} \cdot \boldsymbol{D} \cdot \boldsymbol{P}_{\text{ori}} \tag{10}$$

In this formula, $\boldsymbol{R}_{\text{des}}$ and $\boldsymbol{t}_{\text{des}}$ denote the rotation matrix and translation vector for the destination pose, respectively. Similarly, $\boldsymbol{R}_{\text{ori}}$ and $\boldsymbol{t}_{\text{ori}}$ represent those of the original pose. $\boldsymbol{D}$ is used to convert pixel coordinates $\boldsymbol{P}(x_i, y_i)$ to camera world coordinates (x, y, z). Subsequently, we use the extrinsic matrix to transform the current camera world coordinates into world coordinate systems. Utilizing the destination pose extrinsic matrix transforms the world coordinate system into the target camera coordinate system.

In instances of overlapping points post-rotation, the point with the minimum depth value is retained. We carry out rotational sampling around the initial original pose, rotating along the x, y, and z axes at different angles (5 degrees, 2.5 degrees, and 1.25 degrees) to generate 216 directional poses. We integrate all the rays from the 216 poses, randomly selecting 3,136 rays each time as our spinning around pose.

Helix Rotating Pose. Due to the spiral characteristics of colon folds, the DensiNet module adopts a spiral-shaped sampling trajectory to capture the 3D structure of the folds.

Specifically, we interpolate between the current pose $\boldsymbol{P}_3$ and neighboring pose $\boldsymbol{P}_4$ using the Slerp (Spherical Linear Interpolation) algorithm, which yields a quaternion representing the direction at the intermediate position. The trajectory for the interpolated positions forms a helical path defined by:

$$\begin{cases} \mathrm{x} &= (1-\mathrm{t})\cdot \mathrm{x}_3 + \mathrm{t}\cdot \mathrm{x}_4 + \mathrm{R}\cdot\cos(2\pi\mathrm{t}) \\ \mathrm{y} &= (1-\mathrm{t})\cdot \mathrm{y}_3 + \mathrm{t}\cdot \mathrm{y}_4 + \mathrm{R}\cdot\sin(2\pi\mathrm{t}) \\ \mathrm{z} &= (1-\mathrm{t})\cdot \mathrm{z}_3 + \mathrm{t}\cdot \mathrm{z}_4 + \mathrm{h}\cdot\mathrm{t} \end{cases} \tag{11}$$

Where R should be controlled to be less than the radius of the colon. t represents the position of interpolation, and h denotes the density of interpolation. (x_3, y_3, z_3) and (x_4, y_4, z_4) represent the position of $\boldsymbol{P}_3$ and $\boldsymbol{P}_4$. We execute 400 interpolations and randomly choose one as our helix rotating pose.

Through image warping, we obtain the depth and RGB images in many unseen views, which serve as the pseudo ground truth label. To supervise the colon geometry structure in the rotated view, we compute the discrepancy between these target depths and the depths rendered by the DensiNet under the same poses, using the following loss function:

$$\mathcal{L}_{\text{depth}} = \mathcal{L}_1(\boldsymbol{H}_{\mathrm{d}}, \boldsymbol{D}_3) + \mathcal{L}_1(\boldsymbol{S}_{\mathrm{d}}, \boldsymbol{D}_3) \tag{12}$$

In the above equation, $\mathcal{L}_1$ represents the Smooth L1 Loss [11]. $\boldsymbol{H}_{\mathrm{d}}$ denotes the depth obtained from the helix transformation, $\boldsymbol{S}_{\mathrm{d}}$ is from spin transformation, and $\boldsymbol{D}_3$ corresponds to the depth rendered by the DensiNet for the corresponding transformation method.

The model utilizes these poses to significantly alleviate the sparse viewpoint challenge and explore unseen space around the original pose and the colon wall.

DINO-ViT[6] Vision Transformers (ViT) have been proven to be an effective tool for image texture alignment, possessing the ability to extract valuable texture features [36]. We leverage this tool to address the semantic mismatch between the original and rotated viewpoints - an issue encountered under helix rotating pose and spinning around pose transformations. We aim to maintain stylistic similarity and visual consistency between views after rotation and those from the original viewpoint.

We employ a pre-trained DINO-ViT model [6], which is trained on the ImageNet Datasets [32], for feature extraction. To ensure semantic similarity, we extract tokens to capture the semantic appearance between the original and rotated views. We use the MSE loss to calculate the loss between the extracted features:

$$\mathcal{L}_{\text{ViT}} = \mathcal{L}_{\mathrm{m}}(F_{\text{ViT}}(O_{\mathrm{V}}), F_{\text{ViT}}(R_{\mathrm{V}})) \tag{13}$$

Here, F_{ViT} represents the pre-trained model that we employ to extract semantic information from the RGB of the original views O_{V} and the rendering RGB results in the rotated views R_{V}.

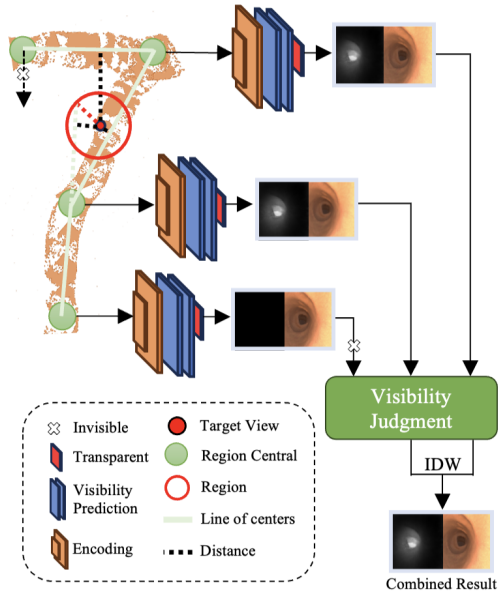

Figure 2. Detailed depiction of our intestinal seamless integration module. The module first evaluates the distance from the line connecting the centers of two blocks to the target view. Blocks exceeding the specified distance threshold, represented by the red area, are filtered out. The remaining blocks undergo visibility prediction, with blocks demonstrating visibility below a certain threshold excluded. The final remaining blocks are seamlessly integrated using Inverse Distance Weighting (IDW), producing our final results.

3.5. Region Integration Module

Filtering Method. To enhance the efficiency of the colon fusion process, we establish two mechanisms for filtering useless blocks. Firstly, as illustrated in Fig. 2, we consider only those blocks within a certain range of the observation points for reliability considerations. Specifically, we calculate the Euclidean distance between the observation points and the line connecting the centers of two adjacent blocks. A block is retained for further processing if this distance is less than 1.5 times the diameter of the colon, ensuring a consistent and reliable selection criterion.

Our second filtering strategy leverages the visibility module, previously introduced in our DensiNet module, to calculate the transparency of this point to the respective block. For each spatial ray, we calculate the transparency metric T_{i} for the ith block in the target view. This transparency metric varies from zero (completely invisible) to one (fully visible). A value approaching one indicates that this 3D point is very close to this DensiNet module, and we can utilize it. Conversely, if the transparency falls be-

low a certain threshold, we exclude that block in the final synthetic process. Our experiments show that the visibility module converges quickly and imposes a negligible computational load due to its small architecture.

Merge Method. To merge adjacent segments after filtering, we employ the Inverse Distance Weighting (IDW) technique proposed by Tancik et al. [34]. We select this method because of its effectiveness in realizing a smooth transition between adjacent segments. This method mitigates the edge jitter that occurred when the merge process relied solely on the closest DensiNet for image rendering.

Specifically, we calculate the distance between the target view P_{t} and the block center, which undergoes a dual filtering process. We determine the merging weight W of each block for interpolation according to the following formula:

$$W \quad = \quad \| \text{ center}, P_{\mathrm{t}} \|^{-\varepsilon} \tag{14}$$

where ε denotes the rendering blend ratio. After calculating the weight W_i for each considerable block i, we normalize it to obtain the weight w_i. Subsequently, we use the following formula to synthesize the final depth and RGB images in the target viewpoint.

$$\begin{cases} Depth & = \sum_0^n w_i * Depth_i \\ RGB & = \sum_0^n w_i * RGB_i \end{cases} \tag{15}$$

where n represents the number of blocks that pass the dual filtering process.

3.6. Overall Objectives

Our final loss function is shown as follows:

$$\mathcal{L}_{\text{all}} = \lambda_1 \mathcal{L}_{depth} + \lambda_2 \mathcal{L}_{ori} + \lambda_3 \mathcal{L}_{ViT} + \lambda_4 \mathcal{L}_{trans} \tag{16}$$

Where λ_1, λ_2, λ_3, and λ_4 represent the weights of different losses, respectively. We aim to balance the numerical scales across different components in determining the weight magnitudes for various loss functions. This operation ensures that each loss term contributes comparably to the overall optimization process. Furthermore, to emphasize the importance of geometric depth supervision, we have intentionally assigned a higher weight to the depth loss component. Thus, we initialize λ_1, λ_2, λ_3, and λ_4 to 8, 1, 10, and 1, respectively.

4. Experiments

4.1. Datasets

To evaluate the performance of our approach, we utilized both synthetic and real-world datasets. Our synthetic dataset comes from the colonoscopy datasets provided by SimCol-to-3D 2022 [31]. We mainly use sequence 1, which is comprehensive in its inclusion of images along with corresponding pose, depth, and intrinsic and extrinsic parameters. For real-world datasets, we primarily employ the C3VD Descending Colon datasets [5] for their applicability in reflecting actual operating conditions.

4.2. Evaluation Metric

We adopt several widely used metrics in comparative view synthesis quality assessment: Peak Signal-To-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS), and the Multi-Scale Structural Similarity Index Measure (MS-SSIM). For LPIPS, we adopt two perceptual metrics based on VGG and AlexNet backbones to ensure a comprehensive evaluation.

4.3. Implementation Details

Our framework is implemented by using PyTorch. All experiments are performed on eight NVIDIA RTX3090 GPUs. The MipNeRF [4] serves as the backbone network. We adopt the Adam optimizer [18] with an initial learning rate of 2e-4, which is progressively reduced during training.

Our synthetic dataset comprises 989 frames; we approximately equally sample one frame for every four as a test set and use the remaining frames as a train set, resulting in 233 test images and 756 train images. The real-world dataset is similarly divided into 35 train images and 19 test images. In our DensiNet, we apply 216 different rotation angles and randomly sample 3,136 rays for training.

4.4. Comparison with State-of-the-Art Methods

We primarily compare our model with several mainstream 3D reconstruction methods, including NeRF [28], MipNeRF [4], FreeNeRF [42], and EndoNeRF [39] on the synthetic dataset [31] and real-world dataset [5]. Before evaluation, we fine-tune the parameters for each scene to ensure a fair comparison.

Qualitative Comparison. As depicted in Fig. 3, our novel view synthesis results on both synthetic and real-world datasets demonstrate significant clarity improvements. Images rendered by NeRF [4], FreeNeRF [42], and EndoNeRF [39] exhibit a noticeable blur, obscuring critical details, such as folds structure, especially within the deeper intestinal regions. Although MipNeRF [4] retains some details, it often learns incorrect geometric shapes. Moreover, the reconstructed depth outcomes from four baselines show significant deviations from the ground truth, potentially misleading in clinical diagnosis.

Our model presents the highest-quality novel view synthesis results, notably in representing folds and the intestinal wall, and provides the clearest rendering results even in deeper areas. It also accurately captures the colon's geometry, which is crucial for precise morphological analysis.

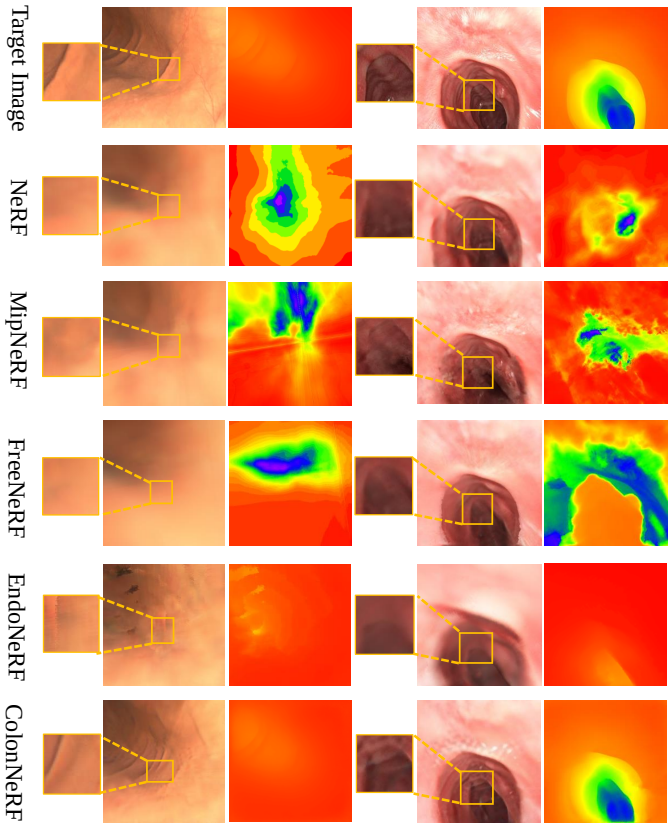

Figure 3. Novel view synthesis results of different methods on the synthetic dataset. The first and third columns display the RGB images, and the second and fourth columns display the corresponding depth images. ColonNeRF consistently outperforms baseline methods with reliably constructed details and a better understanding of geometry.

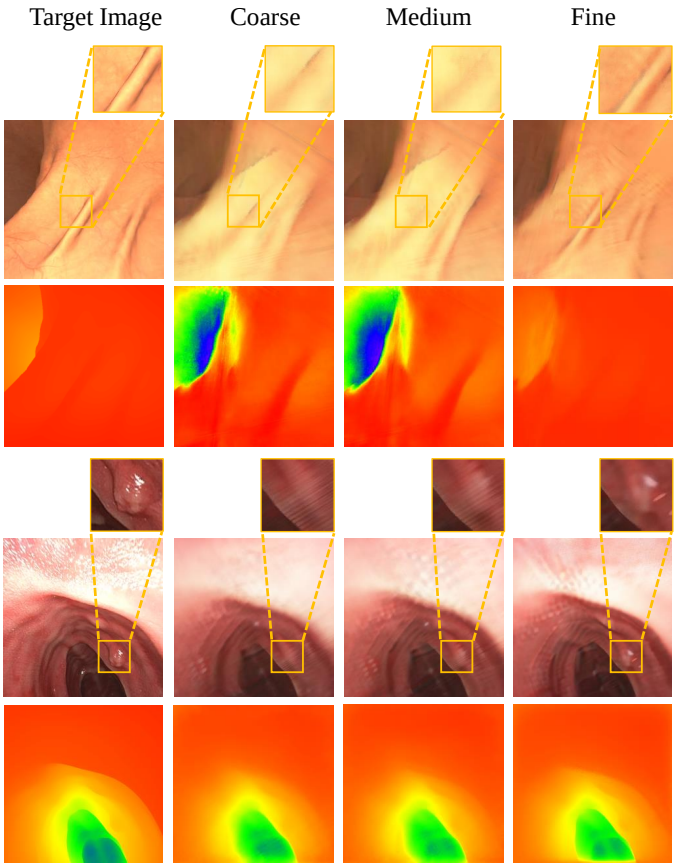

Figure 4. Novel view synthesis from different stages about the multi-level fusion module in the synthetic dataset. With the increase of stage, the model has a more accurate detail reconstruction ability for the scene.

Quantitative Comparison. We mainly compare four evaluation metrics at four baselines and present the results in Tab 1. The labels 'Syn' and 'Real' correspond to outcomes on the SimCol-to-3D [31] and C3VD real-world datasets [5], respectively. Our model demonstrates the highest quantitative performance over all metrics. Specifically, the PSNR metric shows improvements of 2.2% and 3.08% on synthetic and real-world datasets, respectively. With LPIPS-VGG and LPIPS-ALEX metrics on synthetic data, our model outperforms MipNeRF's performance by approximately 21% and 67%. On real-world datasets, we achieve improvements of 2.3% and 6.5% over the best baselines. For the SSIM-MS metric, the improvements are 5% and 3.2% for the synthetic and real-world datasets, respectively. The precise and detailed reconstruction provided by ColonNeRF enables a more accurate morphological analysis of the colon structure. It serves as a dependable reference for clinical assessment and treatment planning, underscoring the superior capability and applicability of our model in medical applications.

Table 1. Quantitative evaluation of our method against four state-of-the-art methods. Compared with four baselines, our model exhibits superior performance in four metrics.

	Datasets	PSNR↑	VGG↓	ALEX↓	MS-SSIM↑
NeRF	Syn	26.10	0.4888	0.4405	0.8266
	Real	25.86	0.4273	0.3745	0.8536
MipNeRF	Syn	24.96	0.4863	0.4367	0.7954
	Real	23.29	0.4142	0.3470	0.7702
FreeNeRF	Syn	24.80	0.5141	0.4815	0.7881
	Real	25.16	0.4096	0.3473	0.8396
EndoNeRF	Syn	21.67	0.4985	0.4378	0.6934
	Real	21.62	0.5077	0.4889	0.7061
ColonNeRF	Syn	**26.70**	**0.3989**	**0.2605**	**0.8373**
	Real	**25.54**	**0.4019**	**0.3242**	**0.8598**

4.5. Ablation Study

Effects of Multi-Level Fusion Module. We explore the efficacy of our multi-level fusion module, investigating both synthetic and real-world datasets, with the results shown in Fig. 4 and Tab. 2. Our analysis involved separate evaluations for each processing stage – coarse, medium, and fine. When the model operates without the multi-level fusion module, meaning it only has the coarse stage, we input

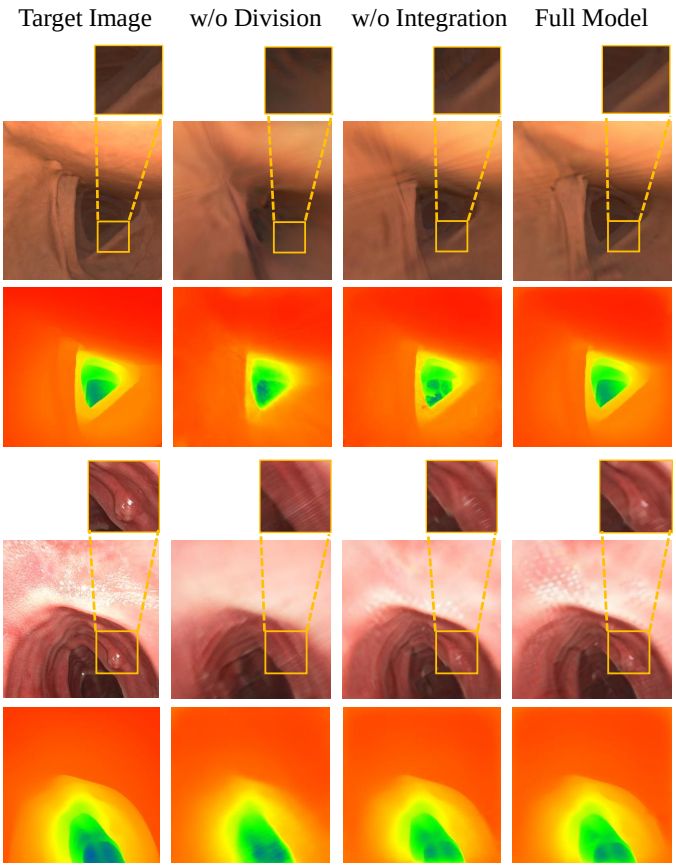

Figure 5. Novel view synthesis without a region division module or integration module. When all the information is inputted into the network simultaneously, numerous lines become distorted, causing a massive decline in quality. With the integration module added, transitional effects at the overpass regions also improve.

Table 2. Ablation Study of multi-level fusion module. The Full Model uses the Fine stage. The table reveals that as the fine-grained data input, the model learns more complex and localized geometry details.

	Datasets	PSNR↑	VGG↓	ALEX↓	MS-SSIM↑
Coarse	Syn	25.28	0.4228	0.2993	0.7986
	Real	24.97	0.4242	0.3393	0.8244
Medium	Syn	25.94	0.4097	0.2770	0.8176
	Real	25.47	0.4143	0.3298	0.8474
Fine	Syn	**26.70**	**0.3989**	**0.2605**	**0.8373**
	Real	**25.96**	**0.4001**	**0.3259**	**0.8676**

the c_1 and σ_1 from the coarse stage directly into the subsequent integration module. As evidenced by the figures, this configuration results in noticeably blurred reconstructions, particularly around the edges.

With the incremental addition of stages, the model progressively reconstructs the colon region from easy to hard and integrates more fine-grained information, resulting in a more comprehensive depiction of detailed information with less noise. Considering efficiency and computational time, we ultimately choose to implement three stages. This module effectively improves the geometric and color results of folds and protrusions.

Table 3. Novel view synthesis without region division or integration module. The results demonstrate that the absence of the division module or the lack of an integration module leads to a decline in quality.

	Datasets	PSNR↑	VGG↓	ALEX↓	MS-SSIM↑
w/o Division	Syn	20.18	0.5883	0.5639	0.6743
	Real	24.49	0.4467	0.3731	0.7944
w/o Integration	Syn	26.62	0.4014	0.2620	0.8344
	Real	25.88	0.4016	0.3288	0.8655
Full Model	Syn	**26.70**	**0.3989**	**0.2605**	**0.8373**
	Real	**25.96**	**0.4001**	**0.3259**	**0.8676**

Effects of Division and Integration Module. We assess the impact of the division and integration modules on our model performance. We present the results in Fig. 5 and Tab. 3. Without the division module, a single block for processing all intestinal data results in noticeable distortions and artifacts. This is because the model is challenging to handle the varied appearance and drastic angle changes in the meandering and convoluted colon. The division module makes each partitioned segment as similar as possible so that our model can better reconstruct the structure of the corresponding region.

Implementing the integration module significantly improves the reconstruction outcomes, especially at transitions between adjacent block regions. This module could combine the understanding of this region from the many blocks to achieve smooth and seamless transitions. The enhanced detail fidelity, accurate geometry, and transition smoothness underscore the importance of the integration module.

Table 4. Ablation study about different views as input. As the number of views increases, with the addition of geometric constraints under various viewpoints, all performance metrics improve, yielding more high-quality outcomes.

	Datasets	PSNR↑	VGG↓	ALEX↓	MS-SSIM↑
1 View	Syn	25.05	0.4407	0.3798	0.8004
	Real	25.47	0.4254	0.4031	0.8523
2 Views	Syn	25.79	0.4086	0.2692	0.8101
	Real	25.77	0.4128	0.3782	0.8533
3 Views	Syn	**26.70**	**0.3989**	**0.2605**	**0.8373**
	Real	**25.96**	**0.4001**	**0.3259**	**0.8676**

Effects of DensiNet Module. We explore the impact of integration inputs from different poses, including the helix rotating pose and the spinning around pose. As depicted in Fig. 6 and Tab. 4, 1 view: original pose as input, 2 views: original pose + helix rotating pose as input, 3 views: original pose + helix rotating pose + spinning around pose as input. With the integration of features from the helix rotating pose, the model demonstrates a significant decrease in blurriness and a marked improvement in understanding the geometric structure of the intestines. Integrating the spinning

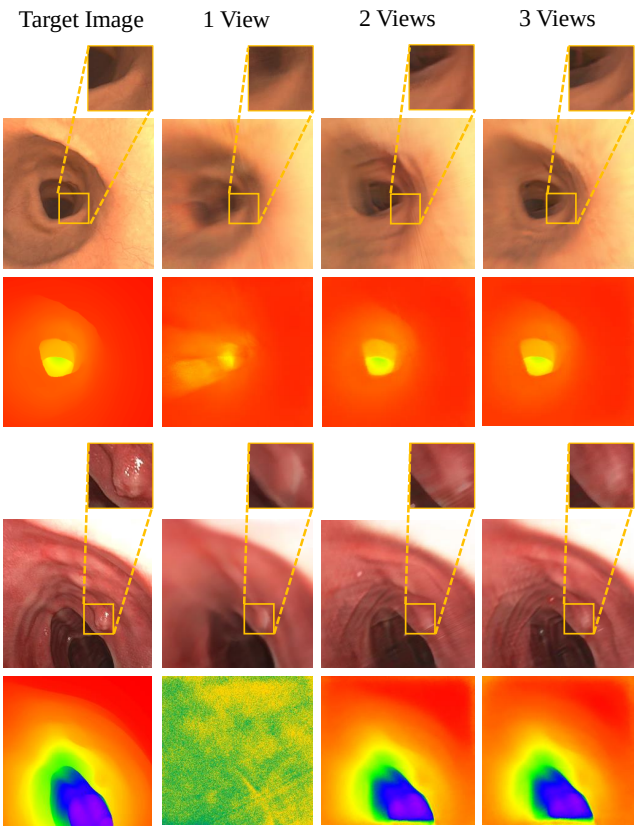

Figure 6. Ablation study about different views as input in synthetic and real-world datasets. The ambiguity significantly diminishes with the increment of input views and enhances the reconstruction quality of the details.

around pose further reduces artifact occurrences, sharpens contours, and enhances depth estimation, resulting in better precision. Our empirical evidence shows that incorporating each new viewpoint provides guidance about semantic consistency and improves the accuracy in depth estimation and the overall clarity of the rendering images.

Effects of Coarse-to-Fine. We conduct an ablation experiment to evaluate the efficacy of the coarse-to-fine strategy in the first block data. This involves contrasting the outcomes of directly inputting fine-grained data at the first stage against a progressive input, transitioning from coarse-grained to fine-grained data. We present the experiment results in Fig. 7 and Tab. 5. By adopting the coarse-to-fine approach, the model learns the simple and complex geometry in a unified framework and progressively models the colon from easy regions to hard regions, thereby yielding reconstructions with improved details.

5. Discussion

The proposed method demonstrates a remarkable ability in synthesizing highly accurate geometry and textures. Notably, our depth results show a significant improvement

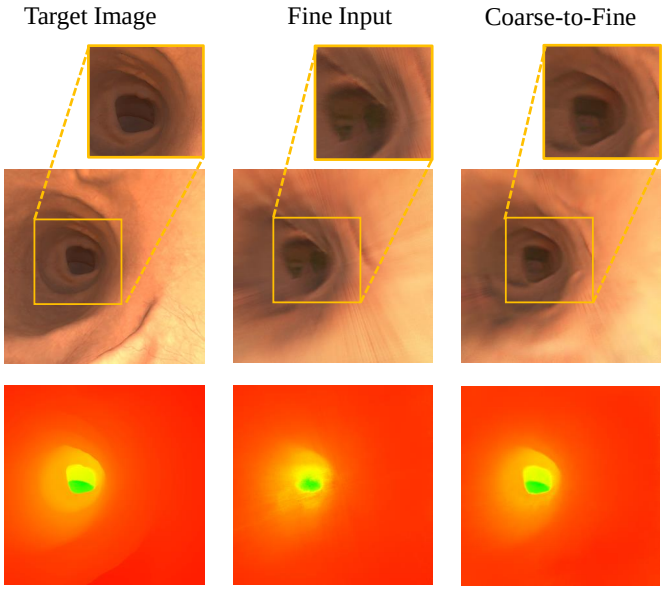

Figure 7. Novel view synthesis from the coarse-to-fine and fine-input ways in the synthetic data. Adopting the coarse-to-fine approach gives the model a better reconstruction effect from easy to hard regions.

Table 5. An ablation study about coarse-to-fine and fine-input which using fine-grained data as input in the first stage.

	PSNR↑	VGG↓	ALEX↓	MS-SSIM↑
Fine-input	26.23	0.3826	0.2488	0.8402
Coarse-to-fine	**27.14**	**0.364**	**0.2214**	**0.8611**

over other methods. A key component contributing to this success is the proposed DensiNet module. This module, employing angular rotation transformations, enables multi-viewpoint co-supervision for geometry estimation, which effectively mitigates modeling difficulties and overcomes the overfitting problem caused by sparse viewpoints.

Although NeRF-based representation provides higher-quality NVS results, it is featured with a long time consumption because of its volume rendering process. A potential solution for improving training efficiency is to propose a more advanced 3D representation strategy that has both modeling flexibility and rendering speed advantages.

6. Conclusions

In this work, we introduced the ColonNeRF, an innovative framework designed for long-sequence colonoscopy reconstruction. To tackle the challenges of such a task, we proposed a region division and integration module to segment long-sequence colons into short blocks, a multi-level fusion module to progressively model the block colons from easy to hard, and a DensiNet module to densify the sampled camera poses under the guidance of semantic consistency. Our extensive testing demonstrates that ColonNeRF outper-

forms four NeRF-based methods in reconstruction quality, proven across both synthetic and real-world environments.

7. Acknowledgement

This project is supported by the National University of Singapore, under the Tier 1 FY2023 Reimagine Research Scheme (RRS).

References

[1] B. Acar et al. Edge displacement field-based classification for improved detection of polyps in ct colonography. *IEEE Transactions on Medical Imaging*, 21(12):1461–1467, Dec 2002. 1

[2] M. Araghi, I. Soerjomataram, M. Jenkins, J. Brierley, E. Morris, F. Bray, and M. Arnold. Global trends in colorectal cancer mortality: projections to the year 2035. *International Journal of Cancer*, 144(12):2992–3000, 2019. 1

[3] John Ashburner and Karl J Friston. Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821, 2000. 2

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3, 4, 5, 7

[5] 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. *arXiv preprint arXiv:2206.08903*, 2022. 7, 8

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[7] Richard J Chen, Taylor L Bobrow, Thomas Athey, Faisal Mahmood, and Nicholas J Durr. Slam endoscopy enhanced by adversarial depth prediction. *arXiv preprint arXiv:1907.00283*, 2019. 1

[8] Tianlong Chen, Peihao Wang, Zhiwen Fan, and Zhangyang Wang. Aug-nerf: Training stronger neural radiance fields with triple-level physically-grounded augmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15191–15202, 2022. 3

[9] Peng Dai, Yinda Zhang, Zhuwen Li, Shuaicheng Liu, and Bing Zeng. Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7830–7839, 2020. 2

[10] D. Freedman et al. Detecting deficient coverage in colonoscopies. *IEEE Transactions on Medical Imaging*, 39(11):3451–3462, Nov 2020. 1

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 6

[12] Ó. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel. Visual slam for handheld monocular endoscope. *IEEE Transactions on Medical Imaging*, 33(1):135–146, Jan 2014. 1

[13] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020. 2

[14] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, and Werner Stuetzle. Mesh optimization. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 19–26, 1993. 2

[15] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022. 3

[16] M. F. Kaminski, J. Regula, E. Kraszewska, M. Polkowski, U. Wojciechowska, J. Didkowska, et al. Quality indicators for colonoscopy and the risk of interval cancer. *New England Journal of Medicine*, 362(19):1795–1803, 2010. 1

[17] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4287–4297, 2021. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[19] AM Leufkens, MGH Van Oijen, FP Vleggaar, and PD Siersema. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy*, pages 470–475, 2012. 1

[20] Marc Levoy et al. Light field rendering (levoy and hanrahan. 1996. 2

[21] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *Advances in Neural Information Processing Systems*, 35:36762–36775, 2022. 3

[22] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion-and view-change human-centric video editing. *arXiv preprint arXiv:2310.10624*, 2023. 3

[23] Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18483–18494, 2023. 3

[24] Xingtong Liu, Zhaoshuo Li, Masaru Ishii, Gregory D Hager, Russell H Taylor, and Mathias Unberath. Sage: slam with appearance and geometry prior for endoscopy. In *2022 International conference on robotics and automation (ICRA)*, pages 5587–5593. IEEE, 2022. 3

[25] Xinyu Liu and Yixuan Yuan. A source-free domain adaptive polyp detection framework with style diversification flow. *IEEE Transactions on Medical Imaging*, 41(7):1897–1908, 2022. 1

[26] Ruibin Ma, Rui Wang, Stephen Pizer, Julian Rosenman, Sarah K McGill, and Jan-Michael Frahm. Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pages 573–582. Springer, 2019. 3

[27] Ruibin Ma, Rui Wang, Yubo Zhang, Stephen Pizer, Sarah K McGill, Julian Rosenman, and Jan-Michael Frahm. Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy. *Medical image analysis*, 72:102100, 2021. 1, 2, 3

[28] 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*, 65(1):99–106, 2021. 2, 3, 5, 7

[29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[30] A. R. Porras, M. Alessandrini, O. Mirea, J. D'hooge, A. F. Frangi, and G. Piella. Integration of multi-plane tissue doppler and b-mode echocardiographic images for left ventricular motion estimation. *IEEE Transactions on Medical Imaging*, 35(1):89–97, Jan 2016. 2

[31] Anita Rau, Binod Bhattarai, Lourdes Agapito, and Danail Stoyanov. Bimodal camera pose prediction for endoscopy. *arXiv preprint arXiv:2204.04968*, 2022. 3, 7, 8

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6

[33] M. Shaban et al. Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE Transactions on Medical Imaging*, 39(7):2395–2405, July 2020. 1

[34] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 3, 7

[35] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 3

[36] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 6

[37] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 3

[38] Shuxian Wang, Yubo Zhang, Sarah K McGill, Julian G Rosenman, Jan-Michael Frahm, Soumyadip Sengupta, and Stephen M Pizer. A surface-normal based neural framework for colonoscopy reconstruction. In *International Conference on Information Processing in Medical Imaging*, pages 797–809. Springer, 2023. 1, 3

[39] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi 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. 1, 2, 3, 7

[40] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *European conference on computer vision*, pages 106–122. Springer, 2022. 3

[41] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 3

[42] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8254–8263, 2023. 7

[43] Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. *arXiv preprint arXiv:2306.13455*, 2023. 3