



OPEN ACCESS

EDITED BY

Hu Cao,
Technical University of Munich, Germany

REVIEWED BY

Yinlong Liu,
City University of Macau, Macao SAR, China
Kexue Fu,
Shandong Academy of Sciences, China

*CORRESPONDENCE

Feng Wang
✉ wf4065335@163.com
Xinrong Chen
✉ chenxinrong@fudan.edu.cn

†These authors have contributed equally to this work

RECEIVED 18 May 2025

ACCEPTED 29 May 2025

PUBLISHED 27 June 2025

CITATION

Ji X, Zhao S, Liu D, Wang F and Chen X (2025)
A robust and effective framework for 3D
scene reconstruction and high-quality
rendering in nasal endoscopy surgery.
Front. Neurobot. 19:1630728.
doi: 10.3389/fnbot.2025.1630728

COPYRIGHT

© 2025 Ji, Zhao, Liu, Wang and Chen. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A robust and effective framework for 3D scene reconstruction and high-quality rendering in nasal endoscopy surgery

Xueqin Ji^{1,2†}, Shuting Zhao^{3,4†}, Di Liu⁵, Feng Wang^{5*} and Xinrong Chen^{3,4*}

¹The Third School of Clinical Medicine, Ningxia Medical University, Yinchuan, China, ²Department of Ultrasound, Peking University First Hospital Ningxia Women and Children's Hospital, Yinchuan, China,

³Fudan University Academy for Engineering and Technology, Shanghai, China, ⁴Shanghai Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention, Shanghai, China,

⁵Department of Hepatobiliary Surgery, General Hospital of Ningxia Medical University, Yinchuan, China

In nasal endoscopic surgery, the narrow nasal cavity restricts the surgical field of view and the manipulation of surgical instruments. Therefore, precise real-time intraoperative navigation, which can provide precise 3D information, plays a crucial role in avoiding critical areas with dense blood vessels and nerves. Although significant progress has been made in endoscopic 3D reconstruction methods, their application in nasal scenarios still faces numerous challenges. On the one hand, there is a lack of high-quality, annotated nasal endoscopy datasets. On the other hand, issues such as motion blur and soft tissue deformations complicate the nasal endoscopy reconstruction process. To tackle these challenges, a series of nasal endoscopy examination videos are collected, and the pose information for each frame is recorded. Additionally, a novel model named Mip-EndoGS is proposed, which integrates 3D Gaussian Splatting for reconstruction and rendering and a diffusion module to reduce image blurring in endoscopic data. Meanwhile, by incorporating an adaptive low-pass filter into the rendering pipeline, the aliasing artifacts (jagged edges) are mitigated, which occur during the rendering process. Extensive quantitative and visual experiments show that the proposed model is capable of reconstructing 3D scenes within the nasal cavity in real-time, thereby offering surgeons more detailed and precise information about the surgical scene. Moreover, the proposed approach holds great potential for integration with AR-based surgical navigation systems to enhance intraoperative guidance.

KEYWORDS

nasal endoscopy, 3D reconstruction, 3D Gaussian Splatting, diffusion model, anti-aliasing

1 Introduction

The demand for endoscopes in transnasal surgery is growing, both for endoscopic examination and endoscopic surgery. For example, according to statistics, rhinosinusitis (RS), an inflammatory disease of the nasal cavity and paranasal sinuses, affects approximately one-six of adults in the United States, resulting in over 30 million diagnoses annually (Wyler and Mallon, 2019; Rosenfeld et al., 2015). Functional endoscopic sinus surgery (FESS), a common method for treating RS, involves inserting a slender endoscope into the nasal cavity to enter the sinus. The endoscope that enters the cavity provides the doctor with a clear field of view, which helps to accurately locate the lesion.

Endoscopy, compared to CT imaging, not only has a lower cost and no radiation, but also better real-time performance, which helps doctors accurately understand the relationship between target lesions and critical anatomical structures (Münzer et al., 2018; Pownell et al., 1997). However, mainstream monocular endoscopy cannot obtain depth information about the internal structure of the nasal cavity, which limits its application in endoscopic examination and endoscopic surgery. Therefore, surface reconstruction from endoscopic sequences enables doctors to obtain 3D information of the internal structure of the nasal cavity, which will better facilitate examination decisions and guide surgical operations.

Structure from Motion (SfM) (Snavely et al., 2006) and Simultaneous Localization and Mapping (SLAM) (Grasa et al., 2013; Mur-Artal et al., 2015) are widely used in depth estimation of endoscopic images, which recover 3D structures by tracking the position of feature points in different images. Widya et al. (2019) investigated how to utilize SfM to overcome the challenge of reconstructing gastric shapes from texture-limited endoscopic images. Leonard et al. (2016) studied an image-enhanced endoscopic navigation method based on the SfM algorithm to improve the accuracy and safety of functional endoscopic sinus surgery. Wang et al. (2020) proposed a bronchoscope enhancement scheme based on visual SLAM, which achieved the reconstruction of feature point models and improved navigation performance; (Mahmoud et al., 2017) successfully stabilized the tracking of endoscope position by combining monocular endoscopy with ORB-SLAM, and successfully repositioned it after tracking loss.

In recent years, neural rendering (Kato et al., 2018; Tewari et al., 2020; Mildenhall et al., 2021) used differentiable rendering and neural networks, surpassing the limited performance of traditional 3D reconstruction. For instance, Wang et al. (2022) utilized dynamic neural radiance fields to represent deformable surgical scenes and explored the potential of neural rendering in 3D reconstruction of surgical scenes. Batlle et al. (2023) introduced LightNeus, which combines neural implicit surface reconstruction technology with photometric models of light sources to achieve 3D reconstruction of the entire colon segment. Chen P. et al. (2024) first utilized Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) to achieve 3D reconstruction of dynamic cystoscopic examination scenes, which can recover scenes under limited perspectives and features, alleviating texture loss problem that traditional algorithms may encounter.

Furthermore, doctors can observe lesion areas from different perspectives through 3D reconstruction using videos obtained from endoscopic examinations, aiding in formulating more precise surgical plans and predicting surgical difficulty and risks. During surgery, the real-time rendering of the 3D scene inside the nasal cavity can be achieved through the posture of the endoscopic camera, providing additional perspective and depth information, which enables doctors to perform cutting, suturing, and other operations more accurately. However, the application of traditional methods in 3D reconstruction of nasal endoscopy has certain limitations. For example, geometry-based reconstruction techniques, such as SfM (Snavely et al., 2006; Widya et al., 2019; Leonard et al., 2016; Schonberger and Frahm, 2016) and SLAM (Grasa et al., 2013; Wang et al., 2020; Mahmoud et al., 2017; Mur-Artal et al., 2015), often struggle to accurately capture feature points

in complex nasal scenes with rich vascular networks and lack of distinct textures, resulting in sparse reconstruction. Additionally, endoscopic images may be affected by lighting effects and lens jitter, leading to image blurring and making reconstruction more complex. The emerging technology based on NeRF (Wang et al., 2022; Batlle et al., 2023; Chen P. et al., 2024) is to use implicit neural representation for volume parameterization of 3D space, which not only is the flexibility poor, but also has slow inference speed, greatly reducing the real-time performance of intraoperative surgery.

Therefore, in this paper, a nasal endoscope reconstruction model, Mip-EndoGS is proposed. Specifically, building upon the foundation of the 3D Gaussian Splatting model (3D-GS) (Kerbl et al., 2023), we employ a diffusion model to alleviate the impact of dynamic blurring in endoscopic images on the reconstruction results. In addition, an adaptive low-pass filter is introduced to reduce aliasing artifacts during the rendering process. We collect a dataset of high-definition surgical videos of nasal examinations performed by professional physicians, recording the spatial position of each frame. Subsequently, we apply the proposed Mip-EndoGS model to this dataset, achieving high-quality and real-time rendering of 3D nasal endoscopic scenes. The main contributions of this paper are as follows.

1. A nasal endoscopy reconstruction model, Mip-EndoGS, is proposed to achieve high-quality 3D reconstruction of nasal endoscopy, which integrates a diffusion module into the 3D Gaussian Splatting framework to remove blur from endoscopic images.
2. An adaptive low-pass filter is embedded into the Gaussian rendering pipeline to overcome aliasing artifacts, which achieves realistic 3D reconstruction of nasal endoscopy scenes.
3. Extensive quantitative and qualitative experiments are conducted to validate the proposed model's effectiveness in reconstructing and rendering nasal endoscopy scenes.

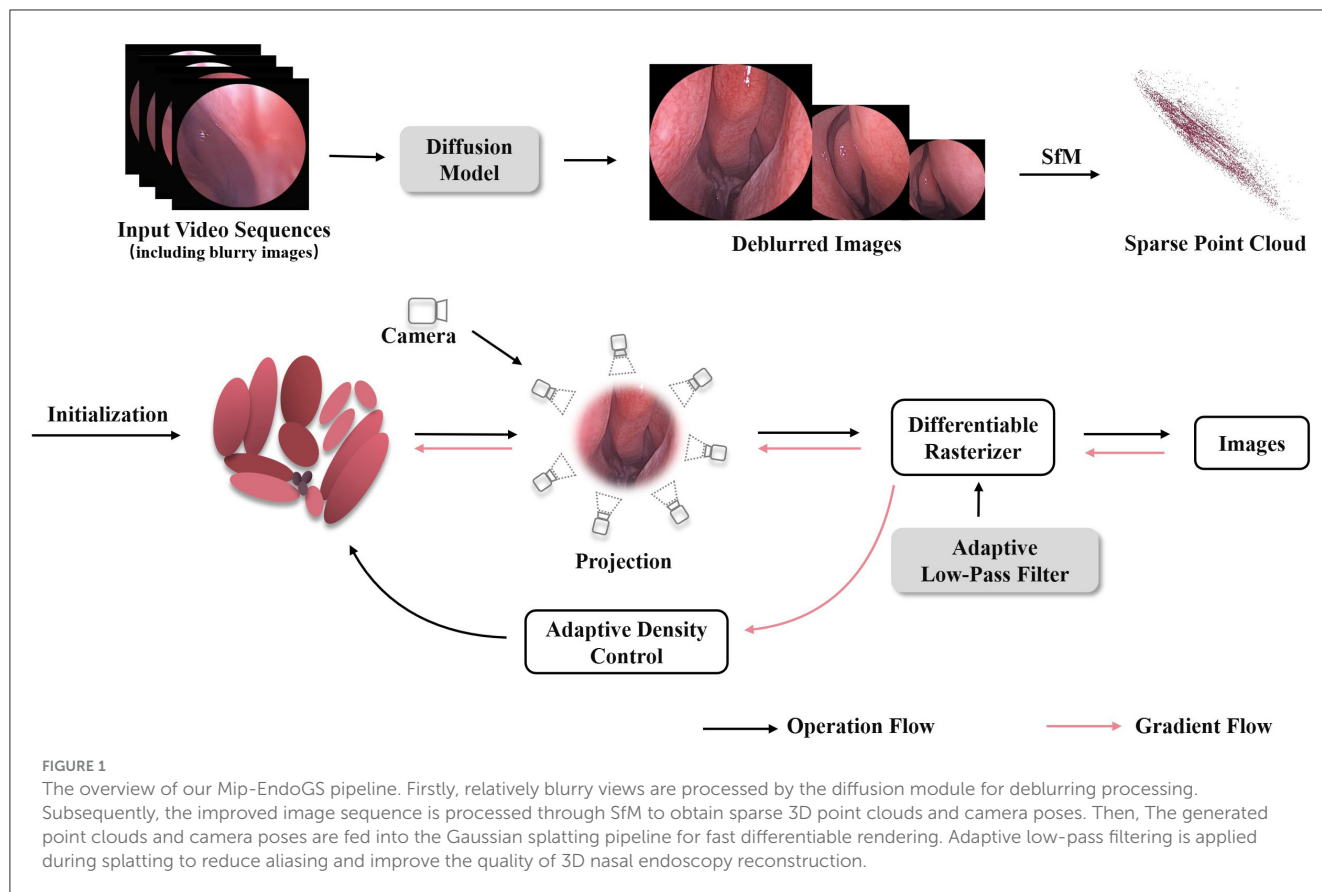
2 Materials and methods

2.1 Nasal endoscopy dataset

The nasal endoscopy dataset, NasED, is constructed by our own. There are 16 subjects with a total of 51 video segments. The data is collected using XION 4K endoscope and NDI optical surgical navigation system. The videos record the process from the inferior and middle nasal meatus to the pharyngeal orifice of the eustachian tube, capturing multi-angle shots of the internal nasal structures, such as the middle and inferior turbinates. Furthermore, the video data is preprocessed into nasal endoscopy examination images with a resolution of 1280 x 720, totaling over 30,000 frames.

2.2 Method architecture

The proposed high fidelity 3D reconstruction and rendering model framework is shown in Figure 1, which comprises two stages, image enhancement based on the diffusion model and 3D-GS differentiable rendering using an adaptive low-pass filter. In the first stage, we uniformly sample several endoscopic views from



the endoscopic video in chronological order and select relatively blurry views as input to the diffusion module (Chen Z. et al., 2024) for deblurring processing. Subsequently, the deblurred views are merged with the original ones to obtain an image-enhanced sequence of endoscopic images. In the second stage, the optimized image sequence is processed through Structure-from-Motion (SfM) (Snavely et al., 2006) algorithms to obtain sparse 3D point clouds and camera poses. These generated point clouds and camera poses are then inputted into the Gaussian splatting pipeline for fast differentiable rasterization rendering (Kerbl et al., 2023). In the splatting rendering process, adaptive low-pass filtering is designed to overcome aliasing issues, thereby achieving high-quality 3D reconstruction of nasal endoscopic scenes.

2.2.1 Image enhancement based on diffusion models

In the process of reconstructing 3D nasal cavity based on endoscopic video, the factors may potentially affect the quality of the images, such as the blurriness caused by the mutual compression of nasal tissues or induced by the dynamic movement of the endoscope. Meanwhile, the potential noise can affect feature extraction between consecutive frames. Therefore, the advanced HI-Diff (Chen Z. et al., 2024) method is employed to denoise the captured nasal endoscopic images, which combines the Transformer based reconstruction module with the traditional diffusion model, and utilizes hierarchical concentration modules (Zamir et al., 2022) to enhance the deblurring process.

The overall framework of HI-Diff deblurring is illustrated in Figure 2. During the training process, given the input blurry image I_{Blur} and its corresponding ground truth image I_{GT} , there are two identical latent encoders (LE) (Rombach et al., 2022) employed to process both images. Specifically, the concatenated form of the blurry image I_{Blur} and its corresponding ground truth image I_{GT} is first fed into one of the latent encoders to extract the prior features v . Simultaneously, the blurry image I_{Blur} is fed into another LE to be mapped to a conditional latent vector p . The specific procedure is as follows:

$$v = f_{LE1}(I_{\text{GT}} \odot I_{\text{Blur}}), \quad (1)$$

$$p = f_{LE2}(I_{\text{Blur}}), \quad (2)$$

Where f_{LE1} and f_{LE2} denote the mappings of the images into high-dimensional space, and $\odot$ represents the concatenation of the two images.

Subsequently, adhering to the procedures outlined in the diffusion model, the prior features are subjected to the addition of random Gaussian noise before being inputted into the denoising network, resulting in v_T . Concurrently, the conditional latent vector p is also fed into the denoising network. This denoising network, conditioned on both inputs, proceeds to predict the ultimate prior features v_1 . The detailed process unfolds as follows:

$$v_T = f_{\text{diffusion}}(v), \quad (3)$$

$$v_1 = f_{\text{denoising}}(v_T, p), \quad (4)$$

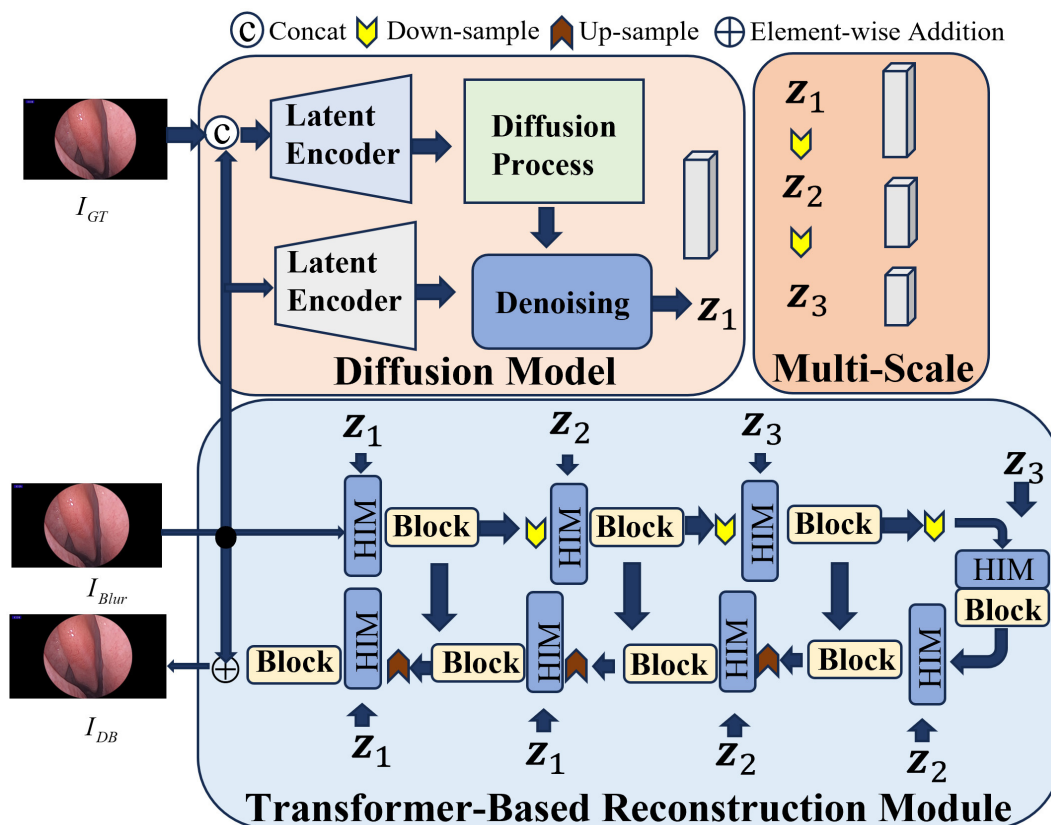


FIGURE 2

The framework of HI-Diff. Blurry views are fed into HI-Diff, which performs deblurring to recover clearer and more detailed structures.

Where $f_{diffusion}$ denotes the process of adding noise to the prior features v , and $f_{denoising}$ represents the neural network. This network takes the vector v_T , which has been augmented with random noise, along with p as inputs, to predict the prior features v_1 .

Moreover, due to the non-uniform blurriness induced by the dynamic motion of the endoscope, relying solely on a single scale of prior features may not adequately accommodate complex blurring scenarios. Hence, to acquire multi-scale prior features capable of adapting to various scales of intermediate features, v_1 is downsampled twice. The specific procedure unfolds as follows:

$$v_2 = f_{down-sample}(v_1), \quad (5)$$

$$v_3 = f_{down-sample}(v_2). \quad (6)$$

For the Transformer-based reconstruction module, given the input blurry image I_{Blur} , the reconstruction module undergoes multiple rounds of upsampling and downsampling before reconstructing the clear image I_{DB} . Furthermore, at each feature extraction stage, a hierarchical concentration module is positioned ahead of both the encoder and decoder, which serves to fuse the intermediate features X_{in} from the Transformer with the multi-scale prior features v_1, v_2, v_3 from the diffusion model through cross-attention fusion. Its purpose is to enhance the deblurring process of the Transformer.

During the testing phase, we replace the ground truth image I_{GT} with randomly generated Gaussian noise. The blurry image I_{Blur}

is then fed into the diffusion model to obtain the prior features v_1 . Subsequently, these prior features are utilized to enhance the blurry image within the Transformer-based reconstruction module, resulting in the generation of high-quality, clear images.

2.2.2 Differentiable rendering through 3D Gaussians Splatting

To achieve fast differentiable rasterization rendering through 3D-GS splatting (Kerbl et al., 2023), the sparse point clouds along with their corresponding camera poses are required, which can be estimated by tracking feature points across multiple images based on Structure-from-Motion (SfM) (Snavely et al., 2006). Based on these point clouds, a set of Gaussian functions using the position mean μ and covariance matrix A is defined. To enhance the representation of the scene, each Gaussian function is equipped with opacity σ and a set of spherical harmonic functions. By introducing this anisotropic 3D Gaussian distribution as a high-quality and unstructured representation of the radiation field, not only can the model compactly represent 3D scenes, but flexible optimization processes are also supported. Specifically, the probability density function of the Gaussian model is as follows (Zwicker et al., 2001b):

$$N(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^T A^{-1}(\mathbf{x}-\mu)}, \quad (7)$$

where A can be decomposed into two more specific components, the quaternion r and the 3D-vector s . Then, these components are transformed into the corresponding rotation and scaling matrices R and S . Therefore, the covariance matrix A can be represented as:

$$A = RSS^T R^T. \quad (8)$$

During the rendering stage, Gaussian elements need to be projected into the rendering space (Zwicker et al., 2001a). Through view transformation, the new covariance matrix in the camera coordinate system can be calculated as follows:

$$A' = JAW^T J^T, \quad (9)$$

where J is the Jacobian matrix approximating the affine transformation of the projection.

Additionally, these Gaussian elements are projected onto the imaging plane according to the observation matrix, and colors are blended based on opacity and depth (Kopanas et al., 2022, 2021). Therefore, the final color $C(p)$ of the p -th pixel can be represented by blending M ordered points overlapping the pixel:

$$C(p) = \sum_{i \in M} T_i \alpha_i c_i, \quad (10)$$

with

$$\alpha_i = \sigma_i e^{-\frac{1}{2}(p-\mu_i)^T A' (p-\mu_i)} \text{ and } T_i = \prod_{j=1}^{i-1} (1 - \alpha_j).$$

T_i is the transmittance, c_i represents the color of the Gaussian element along the direction of the ray, and μ_i denotes the projected 2D UV coordinates of the 3D Gaussians.

Efficient rendering and depth sorting are achieved through a fast tile-based differentiable rasterizer. Additionally, the α -blending technique is introduced to adjust opacity σ and scale parameter S through a sigmoid function, which ensures that the image synthesis maintains higher visual quality.

The rendered scene is compared with the corresponding image to calculate the loss for rapid backpropagation. The loss function consists of $\mathcal{L}_1$ loss and Structural Similarity Index Measure (SSIM), balanced by adjusting the weighting factor λ . It is expressed as follows:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{D-SSIM}. \quad (11)$$

Here, the Stochastic Gradient Descent (SGD) algorithm is utilized to optimize the model parameters iteratively to minimize the loss function (Fridovich-Keil et al., 2022). To further optimize the model, adaptive density control is implemented to adjust the number and density of Gaussian elements for better scene representation. The introduction of transparency threshold ϵ_α and position gradient threshold τ_{pos} is used to control the addition and removal of Gaussian elements. The introduction of this adaptive control allows the method to better adapt to the geometric complexity of nasal endoscopy scenes.

2.2.3 Adaptive low-pass filter

In the rendering process, aliasing is a fundamental issue, as rendered images are usually sampled based on discrete raster grids, inevitably leading to visual artifacts such as jagged edges along object contours and Moiré fringes in textures. A similar phenomenon occurs when splashing elliptical Gaussian, when the scene is reconstructed and rendered at a lower sampling rate.

Previous research (Hu et al., 2023; Yu et al., 2023; Barron et al., 2021) attempts to mitigate aliasing effects generally by prefiltering (Heckbert, 1989; Mueller et al., 1998) and super-sampling techniques (Cook, 1986). For example, the EWA volume reconstruction (Zwicker et al., 2001a) introduces the notion of resampling filters, combining the reconstruction algorithm with a low-pass kernel. Inspired by this method, we employ an anti-aliasing filter to alleviate aliasing artifacts during nasal endoscope rendering. Building upon Equation 10, we further elaborate the rasterization formula:

$$C(p) = \sum_{i=1}^M c_i \sigma_i N'_i(p) \prod_{j=1}^{i-1} (1 - \sigma_j N'_j(p)), \quad (12)$$

Where $N'_i(p)$ represents the projection of the Gaussian distribution onto a two-dimensional plane, closely related to the two-dimensional covariance (Kopanas et al., 2021). And the 2×2 variance matrix A'' can be easily obtained from the 3×3 matrix A' by skipping the third row and column:

$$A' = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} \Leftrightarrow \begin{pmatrix} a & b \\ b & d \end{pmatrix} = A''. \quad (13)$$

Following that, to simulate the diffusion effect occurring during the propagation of light rays, the scale of the 2D covariance is adjusted (Kerbl et al., 2023), for which a positive definite adjustment term is added to the original covariance matrix A'' .

The adjustment term is a scalar multiplied by the unit matrix related to the hyperparameter, by which the scale of the covariance matrix is increased and the diffusion effect of light rays is simulated. Furthermore, in the actual imaging process, the light captured by each pixel accumulates within its surface area, meaning the final image is obtained by integrating the photon energy falling on each pixel (Shirley, 2018). To achieve the actual imaging process more efficiently, the “Adaptive Low-Pass Filter” is proposed as shown in Equation 14, which adapts to different sampling rates and changes in perspective when processing endoscopic images, while maintaining the visual quality of the image.

$$\mathcal{N}^{2D}(x)_{\text{low-pass}} = \sqrt{\frac{|A''|}{|A'' + s\mathbf{I}|}} e^{-\frac{1}{2}(x-\mu)^T (A'' + s\mathbf{I})^{-1} (x-\mu)}. \quad (14)$$

The scale parameter in the adaptive low-pass filter controls the extent of Gaussian smoothing. Intuitively, it simulates the physical diffusion of light across pixel areas due to limited resolution and sampling rates. A larger scale parameter induces stronger anti-aliasing but risks oversmoothing details, while a smaller scale parameter preserves sharpness but may cause jagged edges.

TABLE 1 Quantitative comparison of rendering quality on different video sequences.

Method	H1			H2			H3		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
3D-GS	26.52	0.942	0.114	32.67	0.953	0.137	26.34	0.925	0.156
Mip-EndoGS	27.50	0.936	0.114	35.95	0.971	0.022	30.16	0.934	0.145

The best results are in bold.

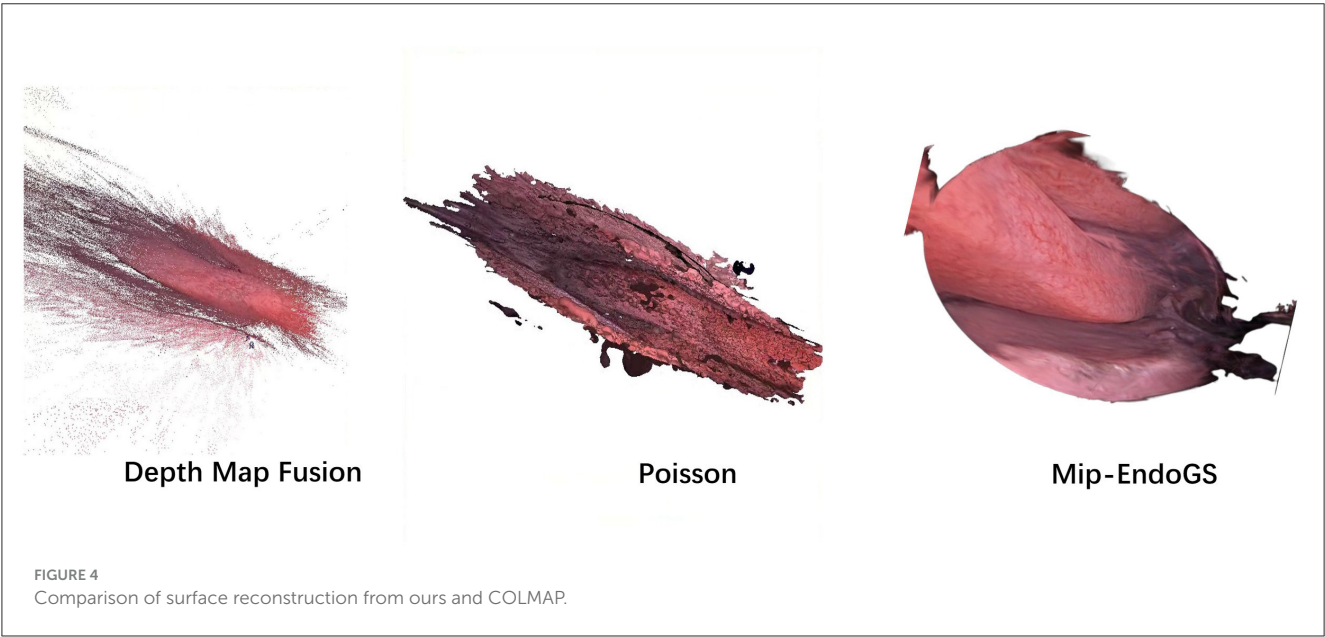
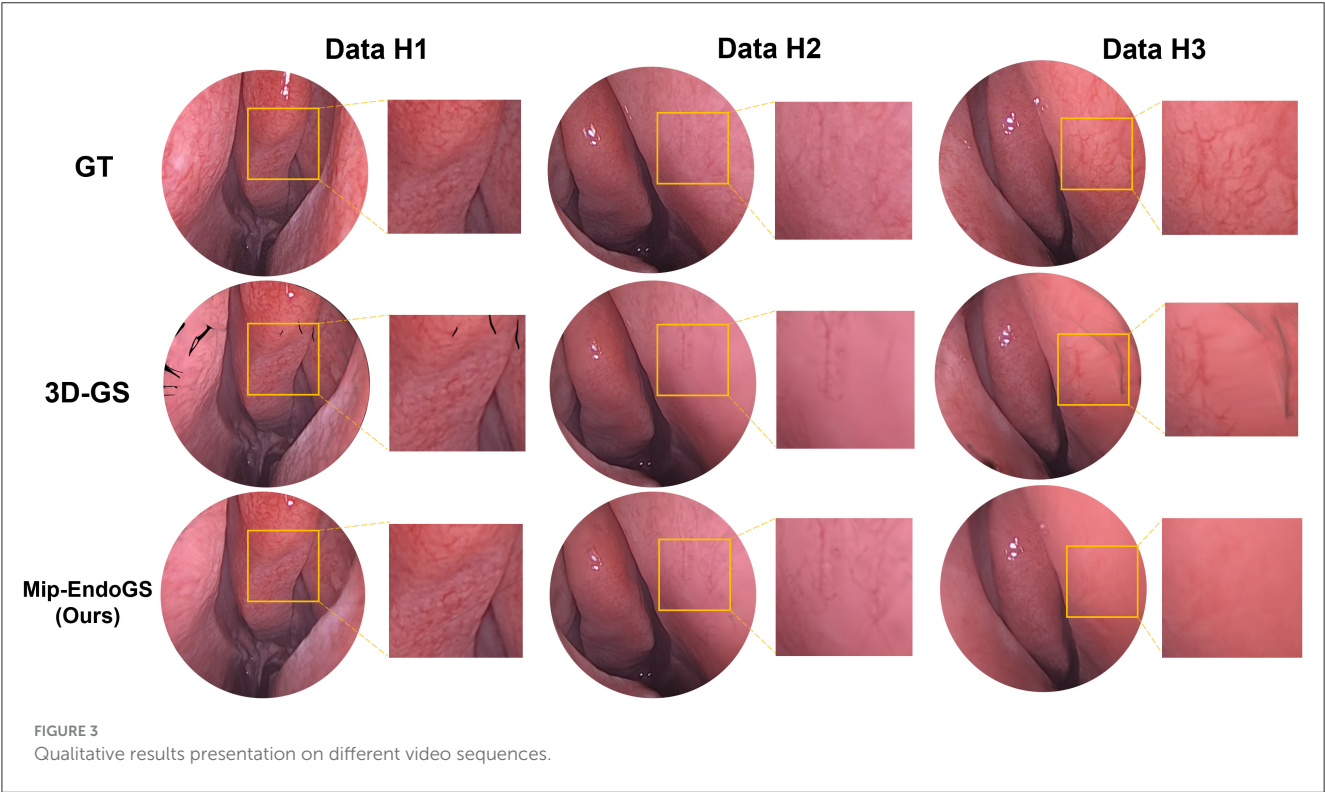


TABLE 2 Quantitative comparison of different outcomes after 6k and 40k iterations.

Method	6k			40k		
	PSNR	Train time	FPS	PSNR	Train time	FPS
3D-GS	20.63	1 m 53 s	112	26.34	12 m 23 s	91
Mip-EndoGS	27.49	3 m 14 s	105	30.16	14m	86

The best results are in bold.

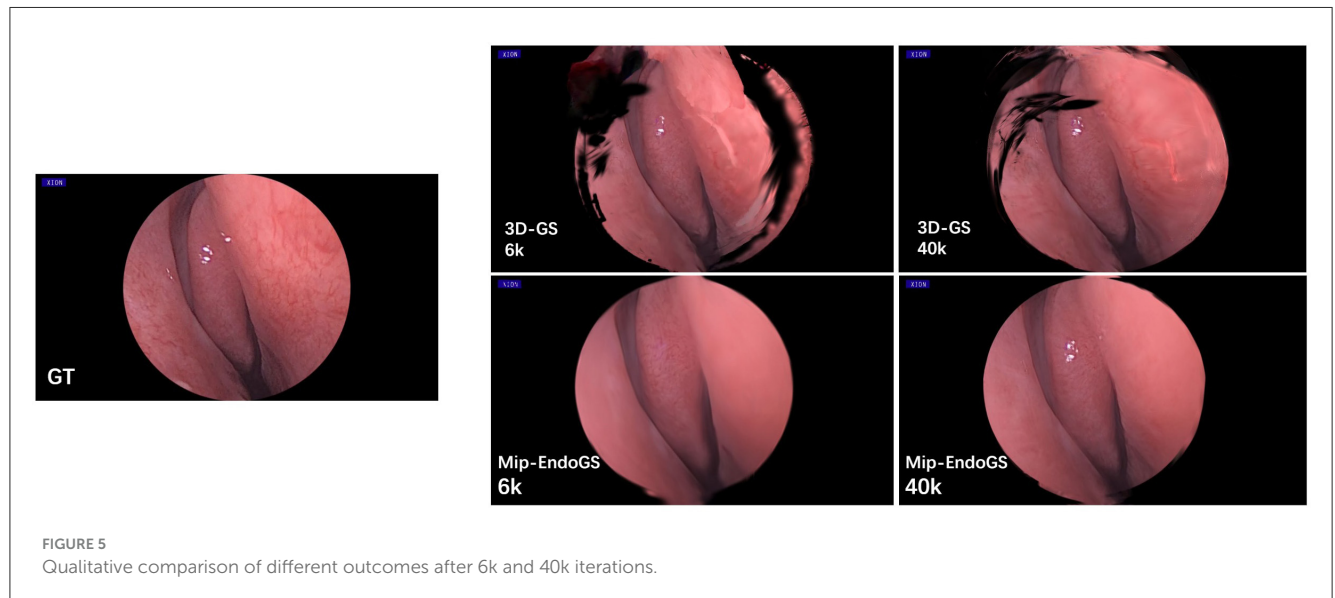


FIGURE 5
Qualitative comparison of different outcomes after 6k and 40k iterations.

3 Results

In this section, the proposed method, Mip-EndoGS, is evaluated based on our nasal endoscopy dataset, NasED. Firstly, the implementation setting of Mip-EndoGS is presented. Then, we provide a detailed introduction of the metrics used in the experiment. Finally, the experimental results are showed, including both quantitative analysis and qualitative analysis.

3.1 Experiments setting

The NasED dataset comprises several monocular nasal endoscopy video sequences, denoted as $\{H_i\}_{i=1}^T$. Here, T represents the number of sequences, and H_i denotes the i -th sequence. Each nasal sequence is divided into several frames, denoted as $\{A_j\}_{j=1}^M$, where M is the total number of frames in the sequence, and j represents the index of the j -th frame. Hence, the i -th sequence and j -th frame's endoscopic view is represented as (H_i, A_j) . From this dataset, we extracted four groups of video sequences: H_1 , H_2 , H_3 , and H_4 . Each group consists of randomly sampled consecutive 100-frame views, totaling 400 frames. Each sequence is split into 90% training data and 10% testing data. These video sequences are captured by a monocular camera, covering the internal structures of the nasal cavity and sinuses.

In the diffusion module, we adhere to experimental settings consistent with HI-Diff and load weights trained on the GoPro

(Nah et al., 2017) synthetic dataset for image denoising. Sparse point clouds and camera poses are obtained through COLMAP (Snavely et al., 2006; Schonberger and Frahm, 2016). The parameters of the Gaussian rendering pipeline (Kerbl et al., 2023) follow the original method settings, except for the changes in the number of iterations. The scale parameter in the adaptive low-pass filter is set to 0.3, and the learning rate is set to $1e-4$. The network is trained on an NVIDIA RTX A6000 device.

3.2 Metric

To conduct a thorough assessment of our experimental results, various methods are employed to evaluate the reconstruction outcomes, primarily comprising quantitative analysis and qualitative assessment through visualization. For quantitative analysis, we utilized several commonly used evaluation metrics, including the Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS).

The computation of SSIM is as follows, which measures the similarity between two images in terms of brightness, contrast, and structure:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (15)$$

TABLE 3 Quantitative comparison of single-scale training and multi-scale testing.

Dataset	Method	PSNR					SSIM					LPIPS				
		Full Res.	1/2 Res.	1/4 Res.	1/8 Res.	Avg.	Full Res.	1/2 Res.	1/4 Res.	1/8 Res.	Avg.	Full Res.	1/2 Res.	1/4 Res.	1/8 Res.	Avg.
H1	3D-GS	26.52	26.50	26.82	29.44	27.32	0.942	0.936	0.910	0.962	0.938	0.114	0.085	0.089	0.052	0.085
	Mip-EndoGS	27.50	27.79	28.66	31.10	28.74	0.936	0.927	0.933	0.954	0.938	0.114	0.084	0.066	0.046	0.076
H4	3D-GS	20.81	20.69	19.87	20.14	20.38	0.883	0.854	0.835	0.812	0.846	0.197	0.203	0.202	0.230	0.208
	Mip-EndoGS	21.33	21.24	21.14	23.73	21.86	0.872	0.832	0.837	0.876	0.854	0.191	0.191	0.180	0.125	0.172

The best results are in bold.

where x and y represent the two images to be compared, μ_x and μ_y denote their mean intensities, σ_x^2 and σ_y^2 represent their variances, σ_{xy} indicates their covariance, and c_1 and c_2 are variables used to stabilize the denominator.

The definition of PSNR is as follows:

$$\text{PSNR} = 20 \cdot \log_{10} \left(\frac{MAX_I}{\sqrt{MSE}} \right), \tag{16}$$

where MAX_I represents the maximum possible pixel value of the image, and MSE is the mean squared error between the reconstructed image and the reference image.

LPIPS employs deep learning models to evaluate the perceptual similarity between images, capturing texture and structural differences crucial for human visual perception:

$$\text{LPIPS}(x, y) = \sum_l w_l \cdot \|\phi_l(x) - \phi_l(y)\|_2, \tag{17}$$

where $\phi_l(x)$ and $\phi_l(y)$ represent the feature maps of images x and y extracted by a pre-trained deep neural network at layer l , and w_l is a learned weight used to emphasize the importance of each layer's contribution to perceptual similarity.

By applying these metrics, we can quantitatively analyze the quality of our image reconstructions.

3.3 Results analysis

3.3.1 Evaluation on full resolution

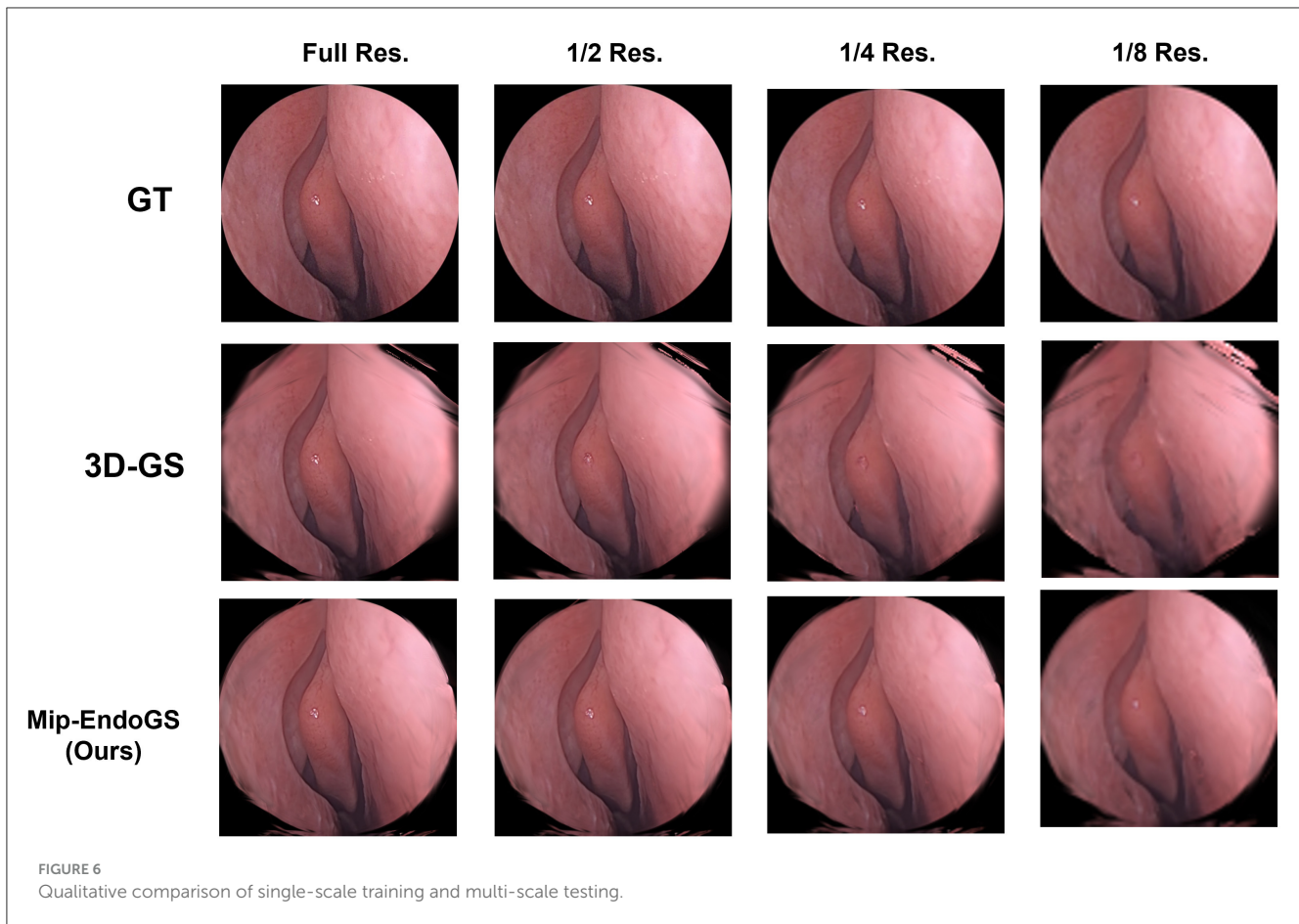
To validate the model's strong generalization capability, we selected sequences from different subjects. The results are presented in Table 1 and Figure 3. Figure 3 illustrates the rendering effects of sequences H_1 , H_2 and H_3 after 40k iterations of model training.

A comparison with the Ground Truth reveals that despite the narrow field of view and lack of texture in nasal endoscopic views, our method renders nasal structures distinctly with clear textures. Compared to the original 3D-GS method, the proposed approach demonstrates higher stability, effectively reducing issues such as significant aliasing, artifacts, and distortions in certain areas observed in the output of 3D-GS. Quantitative evaluation through Table 1 shows notable improvements in the PSNR metrics across all four datasets. Additionally, except for H_1 , the SSIM and LPIPS metrics for the other three datasets also achieve superior results.

3.3.2 Compared with COLMAP

The proposed method, Mip-EndoGS, is compared with the current mainstream reconstruction methods, Depth Map Fusion (Merrell et al., 2007) and the Poisson method (Kazhdan and Hoppe, 2013) in COLMAP. and the visual results are shown in Figure 4.

Data H_1 is utilized in this evaluation. Evidently, the nasal endoscopic scenes reconstruct with Mip-EndoGS exhibit more realistic and smoother features, demonstrating excellent visual outcomes. Apart from comparison with COLMAP, we also attempt reconstruction using methods based on neural radiance fields such as NeRF (Mildenhall et al., 2021) and Neuraludf (Long et al., 2023). However, due to the unique characteristics of nasal structures, these methods all fail.



3.3.3 Evaluation on different iterations

Table 2 and Figure 5 respectively present the quantitative results and visual effects of 3D-GS and Mip-EndoGS at 6k and 40k iterations (evaluated using H_3 data), which shows that our model is capable of capturing the structures within the nasal cavity clearly after 6k iterations, with the PSNR metric significantly outperforming the rendering results of 3D-GS at the same iteration count.

In terms of time, the proposed model only takes around 3 minutes for 6k iterations, less than 1/4 of the time required for 40k iterations. This shorter training time, coupled with clear structural representation, is crucial for real-time surgical navigation. Moreover, although the addition of the diffusion module slightly affects the training time and rendering speed of our model, it still achieves real-time rendering capability.

3.3.4 Evaluation on various resolution

To simulate the reconstruction effects of scenes at low sampling rates, the original data are downsampled to obtain datasets with resolutions reduced to 1/2, 1/4, and 1/8 of the original resolution. We train the model on the original resolution data and render on the downsampled datasets accordingly. The quantitative evaluation is conducted using H_1 and H_4 data (as shown in Table 3), where the proposed method outperforms 3D-GS in rendering quality at lower resolutions. The visual results for H_4 are shown in Figure 6, where the proposed method

produces the higher fidelity imagery without apparent artifacts and aliasing.

4 Discussion

Nasal endoscopic scene reconstruction contributes to a comprehensive understanding of the surgical environment, precise surgical localization, and critical information provision for minimally invasive procedures. However, nasal cavity structures are not only narrow and intricate, but also lack distinctive texture features. Additionally, the influence of endoscopic lighting often makes it challenging to capture nasal cavity structural characteristics. Moreover, the quality of views collected by endoscopy is difficult to guarantee, often resulting in blurriness and contamination.

To address these issues, this paper introduces an advanced nasal endoscopic reconstruction model, Mip-EndoGS, which enables real-time rendering of scenes and synthesis of new viewpoints during surgery by pre-training before surgery. The proposed method consists of two parts, an image enhancement module based on diffusion models and a 3D-GS differentiable rendering pipeline using adaptive low-pass filters. The image enhancement module used in this paper integrates a Transformer-based reconstruction module with traditional diffusion models and employs a hierarchical attention mechanism to enhance the deblurring process of the Transformer, achieving denoising effects on collected nasal endoscopic images. For the differentiable

rendering pipeline based on 3D-GS, we embed an adaptive low-pass filter to overcome aliasing artifacts, which simulates the diffusion effect during light propagation and integrates the photon energy falling on each pixel to adapt to changes in sampling rates and viewpoints.

The proposed method can reconstruct highly realistic nasal endoscopic scenes on the NasED dataset. As shown in the experimental results, the reconstructed nasal structures are distinct with clear textures. Compared to the original 3D-GS, the proposed method demonstrates higher stability, effectively alleviating issues such as aliasing artifacts and distortions during rendering. The high-quality reconstruction results can provide more accurate 3D information, assisting surgeons in diagnosis and reducing surgical risks.

In practice, this task will be combined with motion tracking technology to create a more convenient and intelligent surgical navigation workspace. Additionally, with the development of augmented reality and virtual display technologies, doctors can perform detailed surgical simulations preoperatively and provide real-time three-dimensional views intraoperatively. Such capabilities are particularly valuable in complex or minimally invasive procedures, where accurate spatial perception is critical. These technological advancements can provide doctors with more intuitive and easier-to-use surgical assistance and offer patients higher-quality medical services.

However, certain limitations still exist, such as the occlusions caused by medical instruments and the hands of the surgeon during surgery, as well as deformations of nasal tissues from various angles. These failure cases highlight the need for further optimization in complex surgical environments. To address these challenges, more intelligent surgical planning and navigation technologies are urgently needed.

5 Conclusion

In this work, a novel method, Mip-EndoGS, is proposed to reconstruct the scene of nasal endoscopy. The method combines the diffusion model and 3D Gaussian model, initially employing the diffusion model for deblurring and then achieving high-quality real-time rendering using 3D Gaussian. Additionally, we collect high-definition surgical video datasets from nasal examinations performed by professional doctors and validate the proposed method on this dataset. In the experiment, the proposed method demonstrates superior performance in both quantitative assessment and visual analysis. In the future, we plan not only to expand this dataset but also to further refine the related algorithms.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors upon request.

Ethics statement

The studies involving humans were approved by approved by Eye and ENT Hospital of Fudan University (protocol code 2023188-1 of approval). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

XJ: Data curation, Formal analysis, Methodology, Conceptualization, Writing – original draft. SZ: Software, Writing – original draft, Methodology, Visualization. DL: Investigation, Writing – review & editing, Data curation, Resources, Formal analysis. FW: Writing – review & editing, Data curation, Methodology, Supervision, Project administration. XC: Project administration, Methodology, Supervision, Investigation, Writing – review & editing, Conceptualization.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Ningxia Hui Autonomous Region Key Research and Development Program (Grant No. 2024BEG02018) and the 2022 Medical-Engineering Special Funded Project of the General Hospital of Ningxia Medical University (Grant No. NYZYYG-007).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., and Srinivasan, P. P. (2021). "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 5855–5864.
- Battle, V. M., Montiel, J. M., Fua, P., and Tardós, J. D. (2023). "Lightneus: Neural surface reconstruction in endoscopy using illumination decline," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 502–512.
- Chen, P., Gunderson, N. M., Lewis, A., Speich, J. R., Porter, M. P., and Seibel, E. J. (2024). "Enabling rapid and high-quality 3D scene reconstruction in cystoscopy through neural radiance fields," in *Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling* (New York: SPIE), 350–359.
- Chen, Z., Zhang, Y., Liu, D., Gu, J., Kong, L., Yuan, X., et al. (2024). "Hierarchical integration diffusion model for realistic image deblurring," in *Advances in Neural Information Processing Systems 36* (Cambridge, MA: MIT Press).
- Cook, R. L. (1986). Stochastic sampling in computer graphics. *ACM Trans. Graph.* 5, 51–72. doi: 10.1145/7529.8927
- Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., and Kanazawa, A. (2022). "Plenoxels: Radiance fields without neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 5501–5510.
- Grasa, O. G., Bernal, E., Casado, S., Gil, I., and Montiel, J. (2013). Visual slam for handheld monocular endoscope. *IEEE Trans. Med. Imag.* 33, 135–146. doi: 10.1109/TMI.2013.2282997
- Heckbert, P. S. (1989). *Fundamentals of Texture Mapping and Image Warping*. Available online at: <http://www2.eecs.berkeley.edu/Pubs/TechRpts/1989/5504.html>
- Hu, W., Wang, Y., Ma, L., Yang, B., Gao, L., Liu, X., and Ma, Y. (2023). "Tri-MipRF: Tri-Mip representation for efficient anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Paris: IEEE) 19774–19783.
- Kato, H., Ushiku, Y., and Harada, T. (2018). "Neural 3D mesh renderer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 3907–3916.
- Kazhdan, M., and Hoppe, H. (2013). Screened poisson surface reconstruction. *ACM Trans. Graph.* 32, 1–13. doi: 10.1145/2487228.2487237
- Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 1–14. doi: 10.1145/3592433
- Kopanas, G., Leimkühler, T., Rainer, G., Jambon, C., and Drettakis, G. (2022). Neural point catacaustics for novel-view synthesis of reflections. *ACM Trans. Graph.* 41, 1–15. doi: 10.1145/3550454.3555497
- Kopanas, G., Philip, J., Leimkühler, T., and Drettakis, G. (2021). Point-based neural rendering with per-view optimization. *Comp. Graphics Forum* 40, 29–43. doi: 10.1111/cgf.14339
- Leonard, S., Reiter, A., Sinha, A., Ishii, M., Taylor, R. H., and Hager, G. D. (2016). "Image-based navigation for functional endoscopic sinus surgery using structure from motion," in *Medical Imaging 2016: Image Processing* (New York: SPIE), 235–241.
- Long, X., Lin, C., Liu, L., Liu, Y., Wang, P., Theobalt, C., Komura, T., and Wang, W. (2023). "Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 20834–20843.
- Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J., and Montiel, J. M. M. (2017). "Orb-slam-based endoscope tracking and 3D reconstruction," in *Computer-Assisted and Robotic Endoscopy: Third International Workshop, CARE 2016, Held in Conjunction with MICCAI 2016* (Athens: Springer), 72–83.
- Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.-M., Yang, R., Nistér, D., and Pollefeys, M. (2007). "Real-time visibility-based fusion of depth maps," in *2007 IEEE 11th International Conference on Computer Vision* (Rio de Janeiro: IEEE), 1–8.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 99–106. doi: 10.1145/3503250
- Mueller, K., Moller, T., Swan, J. E., Crawfis, R., Shareef, N., and Yagel, R. (1998). Splatting errors and antialiasing. *IEEE Trans. Visualizat. Comp. Graph.* 4, 178–191. doi: 10.1109/2945.694987
- Münzer, B., Schoeffmann, K., and Böszörményi, L. (2018). Content-based processing and analysis of endoscopic images and videos: a survey. *Multimedia Tools Appl.* 77, 1323–1362. doi: 10.1007/s11042-016-4219-z
- Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D. (2015). Orb-slam: a versatile and accurate monocular slam system. *IEEE Trans. Robot.* 31, 1147–1163. doi: 10.1109/TRO.2015.2463671
- Nah, S., Hyun Kim, T., and Mu Lee, K. (2017). "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 3883–3891.
- Pownell, P. H., Minoli, J. J., and Rohrich, R. J. (1997). Diagnostic nasal endoscopy. *Plastic Reconstruct. Surg.* 99, 1451–1458. doi: 10.1097/00006534-199704001-00042
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 10684–10695.
- Rosenfeld, R. M., Piccirillo, J. F., Chandrasekhar, S. S., Brook, I., Ashok Kumar, K., Kramper, M., et al. (2015). Clinical practice guideline (update): adult sinusitis. *Otolaryngol.-Head Neck Surg.* 152, S1–S39. doi: 10.1177/0194599815572097
- Schonberger, J. L., and Frahm, J.-M. (2016). "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 4104–4113.
- Shirley, P. (2018). *Ray Tracing in One Weekend*. Seattle, WA: Amazon Digital Services LLC, 4.
- Snively, N., Seitz, S. M., and Szeliski, R. (2006). "Photo tourism: exploring photo collections in 3D," in *ACM Siggraph 2006 Papers*, 835–846.
- Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., et al. (2020). State of the art on neural rendering. *Comp. Graphics Forum* 39, 701–727. doi: 10.1111/cgf.14022
- Wang, C., Oda, M., Hayashi, Y., Villard, B., Kitasaka, T., Takabatake, H., et al. (2020). A visual slam-based bronchoscope tracking scheme for bronchoscopic navigation. *Int. J. Comp. Assisted Radiol. Surg.* 15, 1619–1630. doi: 10.1007/s11548-020-02241-9
- Wang, Y., Long, Y., Fan, S. H., and Dou, Q. (2022). "Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer Nature Switzerland), 431–441.
- Widya, A. R., Monno, Y., Okutomi, M., Suzuki, S., Gotoda, T., and Miki, K. (2019). Whole stomach 3D reconstruction and frame localization from monocular endoscope video. *IEEE J. Transl. Eng. Health Med.* 7, 1–10. doi: 10.1109/JTEHM.2019.2946802
- Wyler, B., and Mallon, W. K. (2019). Sinusitis update. *Emerg. Med. Clin.* 37, 41–54. doi: 10.1016/j.emc.2018.09.007
- Yu, Z., Chen, A., Huang, B., Sattler, T., and Geiger, A. (2023). Mip-splatting: Alias-free 3d gaussian splatting. *arXiv [preprint]* arXiv:2311.16493. doi: 10.1109/CVPR52733.2024.01839
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. (2022). "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 5728–5739.
- Zwicker, M., Pfister, H., Van Baar, J., and Gross, M. (2001a). "Ewa volume splatting," in *Proceedings Visualization, 2001* (New York, NY: Association for Computing Machinery), 29–538.
- Zwicker, M., Pfister, H., Van Baar, J., and Gross, M. (2001b). "Surface splatting," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY: Association for Computing Machinery), 371–378.