Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Geometric Initialization to Gaussian Refinement: A Depth-Prior-Guided Framework for Monocular Endoscopic 3D Gaussian Reconstruction

Yizhe Liu
New York University
Jiaqi Huang
Newcastle University
Kejin Zhu
Beihang University
Shixing Ma
Shandong University
Xinzhe Du
Shandong University
Zilin Liang
Beihang University
Cheng Shen
Shenyang Aerospace University
Shuwei Shao
swshao@sdu.edu.cn
Shandong University

Research Article

Keywords: 3D reconstruction, 3D Gaussian Splatting, Collaborative expert depth fusion, Tri-domain pyramid regularization

Posted Date: September 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10827942/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

From Geometric Initialization to Gaussian Refinement: A Depth-Prior-Guided Framework for Monocular Endoscopic 3D Gaussian Reconstruction

Abstract

Monocular endoscopic 3D reconstruction is fundamental to computer-assisted minimally invasive surgery. Recent 3D Gaussian Splatting (3DGS)-based methods enable efficient reconstruction, but their performance strongly depends on the initial geometry, which is commonly obtained through Structure-from-Motion (SfM). In endoscopic scenes, weak textures, specular reflections, and tissue deformation limit stable feature correspondences, often resulting in sparse or incomplete SfM point clouds for Gaussian initialization. Depth back-projection provides an alternative means of constructing dense initial geometry, but estimates from a single depth model may introduce model-specific biases that propagate through subsequent optimization. To alleviate this issue, we propose a collaborative expert depth-guided Gaussian reconstruction framework that improves Gaussian reconstruction from both initialization and refinement perspectives. Specifically, Multi-Model Depth Fusion Initialization robustly fuses depth predictions from multiple models to guide dense Gaussian initialization. During subsequent optimization, Tri-Domain Pyramid Regularization augments spatial-domain supervision with multi-scale pyramid constraints in the wavelet and gradient domains to refine the initialized Gaussian representation. Extensive experiments on two real surgical datasets, SCARED and EndoNeRF, demonstrate that MDGGR achieves state-of-the-art reconstruction quality, with PSNR improvements of 4.4% and 7.8% and LPIPS reductions of 35.7% and 59.2% over prior methods, respectively, particularly in terms of rendering fidelity and fine-grained detail reconstruction.

Keywords: 3D reconstruction, 3D Gaussian Splatting, Collaborative expert depth fusion, Tri-domain pyramid regularization

1 Introduction

Monocular endoscopic 3D reconstruction is fundamental to computer-assisted minimally invasive surgery, supporting intraoperative navigation, surgical scene understanding, and simulation-based training (Brandao et al. 2021; Li et al. 2020; Marcus

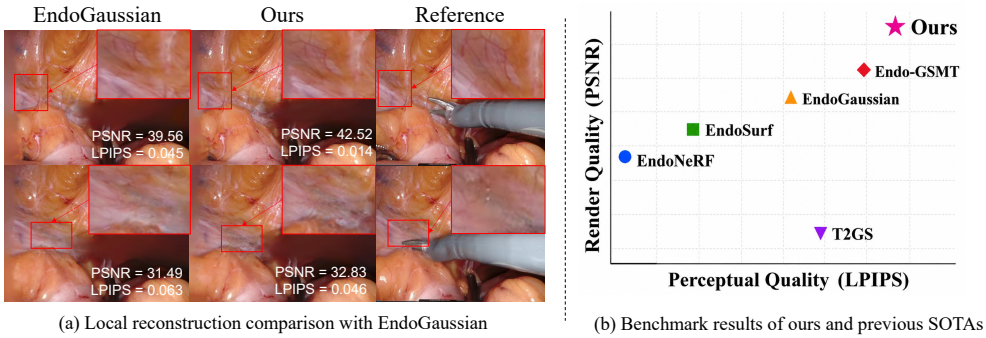
et al. 2014). However, weak textures, specular reflections, occlusions, and tissue deformation make stable inter-frame correspondences difficult to establish, degrading camera estimation and geometric reconstruction (Ozyoruk et al. 2021; Shao et al. 2021, 2022). Early approaches mainly relied on feature-based geometric reconstruction and SLAM pipelines, whose performance remained sensitive to correspondence errors and non-rigid scene dynamics (Chen et al. 2018; Liu et al. 2020, 2022; Zhou and Jagadeesan 2019; Zhou and Jayender 2021). NeRF-based methods subsequently improved continuous scene modeling and novel-view synthesis, yet their volumetric rendering requires prolonged optimization (Mildenhall et al. 2021; Wang et al. 2022; Zha et al. 2023). More recently, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023; Peng et al. 2026) has emerged as an explicit scene representation with substantially lower rendering cost than volumetric neural representations, while relying on an explicit set of Gaussian primitives whose initial geometry affects subsequent reconstruction.

Despite this advantage, the reconstruction quality of 3DGS is closely coupled with the accuracy and completeness of its initial Gaussian geometry. Standard 3DGS pipelines typically initialize Gaussian primitives from Structure-from-Motion (SfM) point clouds, which can become sparse or incomplete in endoscopic videos due to the limited availability of stable inter-frame correspondences. Learning-based monocular depth estimation provides an alternative source of dense geometric information (Bhat et al. 2022; Cui et al. 2024; Shao et al. 2021; Zhou et al. 2026), and recent endoscopic methods have further explored learned depth or dense geometric predictions as scene priors (Sheikh Zeinoddin et al. 2024, 2025). By back-projecting predicted depth maps into three-dimensional space, dense point sets can be constructed for Gaussian initialization.

However, monocular depth estimators trained with different data, supervision strategies, and architectures can exhibit model-specific deviations in scale, local structure, and boundary localization (Bhat et al. 2022; Cui et al. 2024; Shao et al. 2021; Sheikh Zeinoddin et al. 2024, 2025). When a prediction from a single estimator is back-projected, these deviations may translate into geometric discrepancies in the initial point positions and Gaussian centers, where they can persist or propagate during subsequent optimization (Kerbl et al. 2023; Liu et al. 2024). The distinct prediction characteristics of different estimators therefore provide complementary geometric information, motivating multi-model depth fusion for Gaussian initialization.

While improved initialization provides a better geometric starting point, the subsequent Gaussian optimization still requires effective supervision. Recent studies have extended Gaussian-based endoscopic reconstruction toward efficient optimization (Liu et al. 2024), deformable tissue modeling and 4D representation (Huang et al. 2024; Zhu et al. 2024), and sequential reconstruction (Zhang et al. 2026). Despite these advances in representation and modeling capability, their reconstruction objectives are still primarily formulated in the spatial domain, using photometric and depth supervision together with deformation, smoothness, or surface regularization (Huang et al. 2024; Liu et al. 2024; Zhu et al. 2024). Although effective for overall appearance and geometry reconstruction, these objectives do not explicitly constrain localized high-frequency discrepancies or structural boundaries. This limitation is particularly relevant in endoscopic scenes, where weak texture and tissue deformation increase the

ambiguity of pixel-wise supervision (Ozyoruk et al. 2021; Shao et al. 2021; Stoyanov 2012), while specular reflections and illumination variation disturb local appearance and edge reconstruction (Bidokh and Hassanpour 2023; Song et al. 2022). We therefore introduce a tri-domain pyramid regularization strategy that combines spatial-domain supervision with multi-scale wavelet- and gradient-domain constraints to improve local structural and high-frequency consistency during Gaussian refinement.



(a) Local reconstruction comparison with EndoGaussian

(b) Benchmark results of ours and previous SOTAs

Fig. 1 Visual comparison and benchmark results on the EndoNeRF dataset. (a) Representative reconstruction comparison between EndoGaussian and our method. (b) PSNR–LPIPS comparison with existing methods.

In this work, we propose a depth-prior-guided Gaussian reconstruction framework. Collaborative Expert Depth Fusion first constructs a fused depth prior by aligning predictions from multiple depth experts through scale-shift fitting and integrating the aligned predictions through residual-dependent weighting. The fused prior is then back-projected through Depth-Guided Gaussian Initialization to establish a dense Gaussian representation. Tri-Domain Pyramid Regularization subsequently imposes complementary spatial-domain, wavelet-domain, and gradient-domain constraints during Gaussian refinement. Together, Collaborative Expert Depth Fusion and Tri-Domain Pyramid Regularization improve reconstruction at complementary phases: the former provides the fused depth prior used to establish the initial geometry, while the latter constrains its subsequent refinement. Fig. 1 provides an intuitive comparison of the reconstruction quality achieved by the proposed method and existing approaches.

The main contributions of this work are summarized as follows:

- We propose Collaborative Expert Depth Fusion, which aligns predictions from multiple depth experts through scale-shift fitting and robustly integrates the aligned predictions to construct a fused depth prior for Gaussian initialization.
- We introduce Tri-Domain Pyramid Regularization, which augments spatial-domain Gaussian optimization with multi-scale wavelet- and gradient-domain constraints to preserve anatomical details and suppress localized artifacts.
- Extensive experiments on the SCARED and EndoNeRF datasets demonstrate that the proposed framework achieves superior reconstruction quality, with

PSNR/SSIM/LPIPS of 40.497/0.976/0.018 on EndoNeRF and 29.719/0.865/0.089 on SCARED.

2 Related Work

2.1 Traditional and Neural Endoscopic Reconstruction

Endoscopic 3D reconstruction has been widely studied for computer-assisted minimally invasive surgery. Traditional methods mainly rely on geometry-based pipelines, including Structure-from-Motion, SLAM-based reconstruction, stereo matching, and multi-view reconstruction (Chen et al. 2018; Grasa et al. 2013; Schonberger and Frahm 2016; Seitz et al. 2006; Zhou and Jagadeesan 2019; Zhou and Jayender 2021). However, these methods require sufficient and geometrically consistent feature correspondences, as well as cross-view photometric consistency, which are difficult to maintain in endoscopic scenes due to weak textures, specular highlights, occlusions, and tissue deformation (Ozyoruk et al. 2021; Shao et al. 2021; Stoyanov 2012; Zhong et al. 2025). Neural rendering methods such as NeRF (Mildenhall et al. 2021) and endoscopy-specific variants including EndoNeRF (Wang et al. 2022) and EndoSurf (Zha et al. 2023) improve reconstruction quality by modeling surgical scenes with continuous implicit representations. Nevertheless, their volumetric rendering process remains computationally expensive and may still produce over-smoothed anatomical structures. Endoscopic video restoration studies also show that recovering fine visual details from degraded endoscopic videos is important for improving downstream perception quality, which is closely related to the high-frequency artifact suppression problem considered in this work (Song et al. 2022). In contrast, our method adopts an explicit Gaussian representation initialized from a collaboratively fused multi-expert depth prior and refined with multi-domain constraints.

2.2 3D Gaussian Splatting in Endoscopic Reconstruction

3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) provides an efficient alternative to implicit neural rendering by representing scenes with explicit anisotropic Gaussian primitives. Recent studies have extended 3DGS to deformable and endoscopic reconstruction. EndoGaussian (Liu et al. 2024) introduces Gaussian splatting for deformable endoscopic scenes, while EndoGS (Zhu et al. 2024), Endo-4DGS (Huang et al. 2024), Endo-4DGX (Huang et al. 2025), and SG-3DGS (Zhang et al. 2026) further improve deformation modeling, surface consistency, illumination robustness, or sequential reconstruction. Despite these advances, Gaussian-based endoscopic reconstruction remains sensitive to initialization quality and refinement stability. In monocular settings, sparse SfM initialization often yields incomplete and geometrically inconsistent point sets, while directly using noisy depth priors may cause geometric distortions or convergence instability during Gaussian optimization.

In addition to initialization, stable Gaussian refinement also requires effective regularization. Existing Gaussian-based pipelines mainly rely on spatial-domain rendering supervision, depth consistency, and smoothness constraints (Mahendran and Vedaldi 2015). These constraints preserve overall appearance and suppress local noise, but

they are often insufficient for preserving sharp anatomical boundaries and suppressing localized high-frequency artifacts in weakly textured or reflective regions. To improve refinement stability, our framework augments the original spatial-domain optimization with wavelet-domain and gradient-domain pyramid constraints, forming Tri-Domain Pyramid Regularization for more stable and detail-preserving Gaussian refinement.

2.3 Monocular Depth Priors for Endoscopic Reconstruction

Monocular depth estimation provides important geometric guidance for endoscopic reconstruction. Self-supervised and endoscopy-oriented methods have improved depth prediction robustness in surgical scenes by exploiting temporal consistency, structural constraints, domain-specific adaptation (Cui et al. 2024; Sheikh Zeinoddin et al. 2024, 2025) or foundation-model priors (Bhat et al. 2022; Shao et al. 2021, 2026; Zhu et al. 2026). Recent studies have further explored CLIP-guided depth learning, confidence-guided dual-branch estimation, and multi-scale pyramid-based depth estimation for endoscopic or laparoscopic scenarios (Chong and Yang 2025; Xiang et al. 2026; Yang et al. 2024). Despite these advances, depth priors derived from a single model may still suffer from model-specific bias, local structural inconsistency, and scale-shift mismatch in textureless, reflective, or occluded regions. Directly using such priors for Gaussian initialization can therefore introduce unstable geometric support. To reduce expert-specific deviations in the depth prior, our framework aligns complementary predictions from multiple depth experts and applies Collaborative Expert Depth Fusion to construct a more structurally consistent depth prior for Gaussian initialization.

3 Method

3.1 Overall Framework

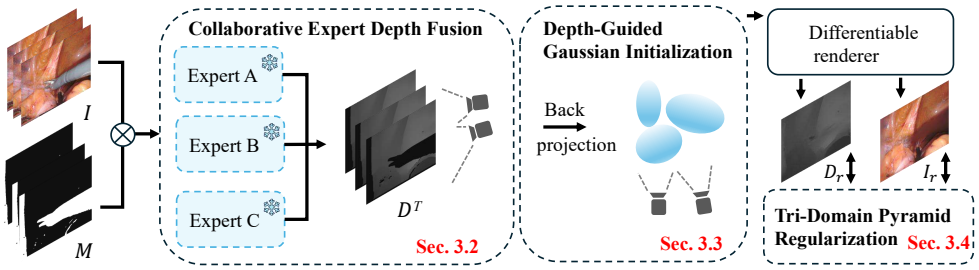


Fig. 2 Overview of the proposed monocular endoscopic 3D reconstruction framework. The framework consists of two complementary stages: Collaborative Expert Depth Fusion constructs a fused depth prior from multiple monocular depth experts for dense Gaussian initialization, while Tri-Domain Pyramid Regularization refines the initialized Gaussian representation through multi-scale constraints in the spatial, wavelet, and gradient domains.

The overall pipeline of the proposed framework is illustrated in Fig. 2. Given a monocular endoscopic image sequence, the framework constructs a Gaussian representation through depth-prior-guided initialization and multi-domain-regularized refinement. Accordingly, the framework contains two proposed components, Collaborative Expert Depth Fusion and Tri-Domain Pyramid Regularization, connected by depth-guided back-projection for Gaussian initialization.

Collaborative Expert Depth Fusion aligns and integrates predictions from multiple monocular depth experts to construct a fused depth prior. Depth-Guided Gaussian Initialization then back-projects this prior into a dense 3D point set used to initialize the Gaussian primitives instead of the sparse point cloud typically obtained through SfM. Tri-Domain Pyramid Regularization subsequently augments spatial-domain rendering supervision with multi-scale wavelet- and gradient-domain constraints during Gaussian refinement. The first two parts construct the initial Gaussian representation, whereas Tri-Domain Pyramid Regularization constrains its subsequent optimization.

3.2 Collaborative Expert Depth Fusion

Collaborative Expert Depth Fusion constructs a geometrically consistent depth prior by aligning predictions from multiple monocular depth estimators and suppressing inconsistent estimates during fusion. The overall process, including scale-shift alignment, median initialization, and iteratively reweighted least-squares fusion, is illustrated in Fig. 3.

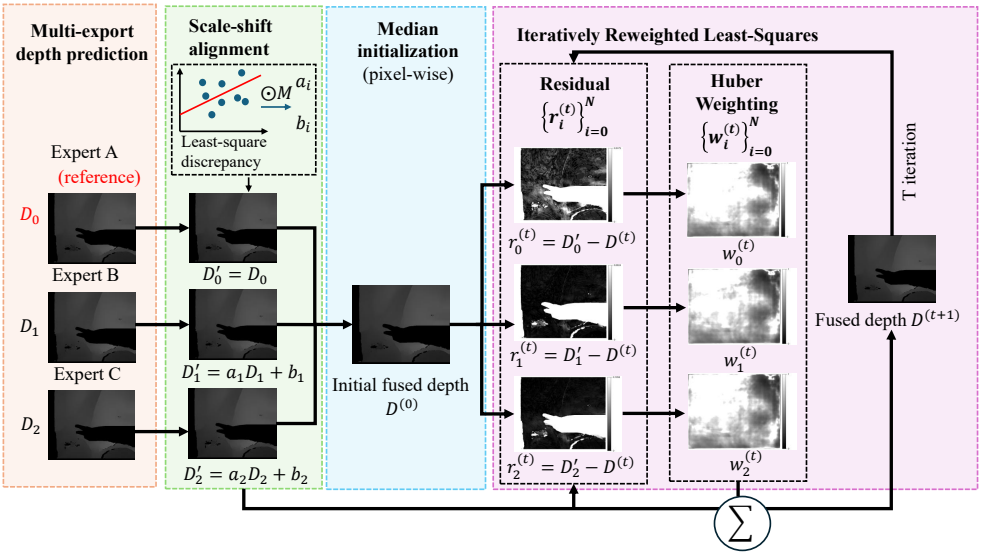


Fig. 3 Mechanism of the proposed Collaborative Expert Depth Fusion. Aligned depth estimates are iteratively fused using expert-level fusion weights and pixel-wise Huber weights to obtain a structurally coherent fused depth for Gaussian initialization.

Endoscopic reconstruction based on depth back-projection initialization is highly dependent on the quality of the externally provided depth priors. Predictions from a single depth estimator may contain model-specific biases and local structural inconsistencies, particularly in weakly textured, reflective, or occluded regions. To reduce reliance on any individual estimator, we employ multiple depth models with different architectures and training paradigms. Their predicted depth maps are represented by $\{\boldsymbol{D}_i\}_{i=0}^{N}$, where $N = 2$ in our implementation.

Because monocular depth estimators may produce predictions with inconsistent scales and shifts, directly combining their outputs may introduce geometric inconsistencies. We therefore use $\boldsymbol{D}_0$ to define a common scale-shift coordinate space and transform each remaining prediction through an affine operation:

$$\boldsymbol{D}_i' = a_i \boldsymbol{D}_i + b_i. \tag{1}$$

where $\boldsymbol{D}_i$ is the depth map generated by the i-th estimator, while a_i and b_i correspond to its scale and shift parameters, respectively. The transformed depth prediction is represented by $\boldsymbol{D}_i'$.

To estimate the transformation parameters over valid image regions, a_i and b_i are obtained by minimizing the masked least-squares discrepancy between the transformed prediction and $\boldsymbol{D}_0$:

$$(a_i, b_i) = \arg\min_{a,b} \|\boldsymbol{M} \odot (a\boldsymbol{D}_i + b - \boldsymbol{D}_0)\|_2^2 . \tag{2}$$

The mask $\boldsymbol{M}$ covers the valid pixels used for cross-model alignment, and $\odot$ represents element-wise multiplication. For the reference prediction, we set $\boldsymbol{D}_0' = \boldsymbol{D}_0$.

The prediction $\boldsymbol{D}_0$ is used to establish the common affine coordinate space required for alignment rather than to introduce a primary–auxiliary hierarchy among the depth estimators. After alignment, all predictions are incorporated into the same robust fusion procedure. This operation reduces relative scale-shift discrepancies across the models without recovering an absolute metric depth scale.

To improve robustness against outliers and local cross-model inconsistencies, the fusion process is initialized using the pixel-wise median of the aligned predictions $\{\boldsymbol{D}_i'\}_{i=0}^{N}$. The resulting initial estimate is written as $\boldsymbol{D}^{(0)}$. Compared with direct averaging, median initialization reduces the influence of individual predictions that deviate substantially from the remaining models.

An Iteratively Reweighted Least Squares procedure is then applied to refine the fused depth prior. To measure the discrepancy between each aligned prediction and the current fused estimate, the residual map at iteration t is computed as

$$\boldsymbol{r}_i^{(t)} = \boldsymbol{D}_i' - \boldsymbol{D}^{(t)}. \tag{3}$$

In this expression, $\boldsymbol{D}^{(t)}$ is the fused depth estimate obtained at iteration t, and $\boldsymbol{r}_i^{(t)}$ represents the corresponding residual map of the i-th estimator.

To reduce the influence of locally inconsistent predictions, each residual map is converted into a pixel-wise Huber weight:

$$\boldsymbol{w}_i^{(t)} = \begin{cases} 1, & |\boldsymbol{r}_i^{(t)}| \le \delta, \\ \dfrac{\delta}{|\boldsymbol{r}_i^{(t)}|}, & |\boldsymbol{r}_i^{(t)}| > \delta, \end{cases} \tag{4}$$

where δ is the Huber threshold. The weight map $\boldsymbol{w}_i^{(t)}$ assigns larger weights to predictions that are locally consistent with the current fused estimate and adaptively reduces the weights of predictions with large residuals.

In addition to the residual-dependent Huber weights, a scalar model-level weight α_i is assigned to each depth estimator to control its overall contribution to the fusion process. The model-level weights regulate the global contributions of the estimators, whereas $\boldsymbol{w}_i^{(t)}$ adjusts their contributions independently across different pixel locations.

To jointly incorporate model-level weighting and pixel-wise residual consistency, the fused depth prior is updated as

$$\boldsymbol{D}^{(t+1)} = \frac{\sum_{i=0}^{N} \alpha_i \boldsymbol{w}_i^{(t)} \odot \boldsymbol{D}_i'}{\sum_{i=0}^{N} \alpha_i \boldsymbol{w}_i^{(t)}}. \tag{5}$$

The summation is performed across the depth estimators, while the division is evaluated element-wise. The scalar weights α_i determine the global contribution of each estimator to the fused estimate, while the Huber weighting mechanism suppresses spatially localized inconsistencies during each iteration. Differences in α_i specify different global fusion contributions and do not imply a primary–auxiliary model structure.

After T iterations, the resulting fused depth prior $\boldsymbol{D}^{(T)}$ provides geometric guidance for subsequent Gaussian initialization rather than being interpreted as an absolute metric depth estimate. By reducing reliance on any single expert and attenuating locally inconsistent predictions, the proposed collaborative expert fusion strategy produces a more stable and structurally coherent depth prior for depth-guided Gaussian initialization.

3.3 Depth-Guided Gaussian Initialization

Monocular endoscopic scenes present specific challenges, including weak textures, illumination variations, and non-rigid tissue deformation, which often undermine the effectiveness of SfM. To overcome this limitation, we leverage fused depth priors to back-project pixels from the training frames into 3D space, thereby constructing a dense point cloud for Gaussian initialization.

Given the fused depth map $\boldsymbol{D}^{(T)}$, we back-project pixels in the 2D image plane into the camera coordinate system. Specifically, the 3D point corresponding to pixel $\mathbf{p} = (u, v)$ is computed as

$$\mathbf{X}(u, v) = d\,\mathbf{K}^{-1}\tilde{\mathbf{p}}, \tag{6}$$

where $d = \boldsymbol{D}^{(T)}(u, v)$ is the depth value at pixel $\mathbf{p} = (u, v)$. The matrix $\mathbf{K}$ represents the camera intrinsic matrix, while $\tilde{\mathbf{p}}$ is the homogeneous coordinate of the corresponding pixel. The resulting dense point set is then used to initialize the Gaussian means and associated attributes.

3.4 Tri-Domain Pyramid Regularization

Existing methods incorporate inverse-depth supervision into spatial-domain objectives to constrain Gaussian geometry(Liu et al. 2024; Yang et al. 2024). Nevertheless, such supervision is highly dependent on the estimated depth and does not explicitly penalize high-frequency appearance or boundary discrepancies. We therefore introduce a tri-domain pyramid regularization mechanism, which extends spatial-domain supervision with multi-scale wavelet domain and gradient domain constraints to constrain high-frequency responses and anatomical contours, respectively, enabling finer refinement of the Gaussian representation. The overall mechanism is illustrated in Fig. 4.

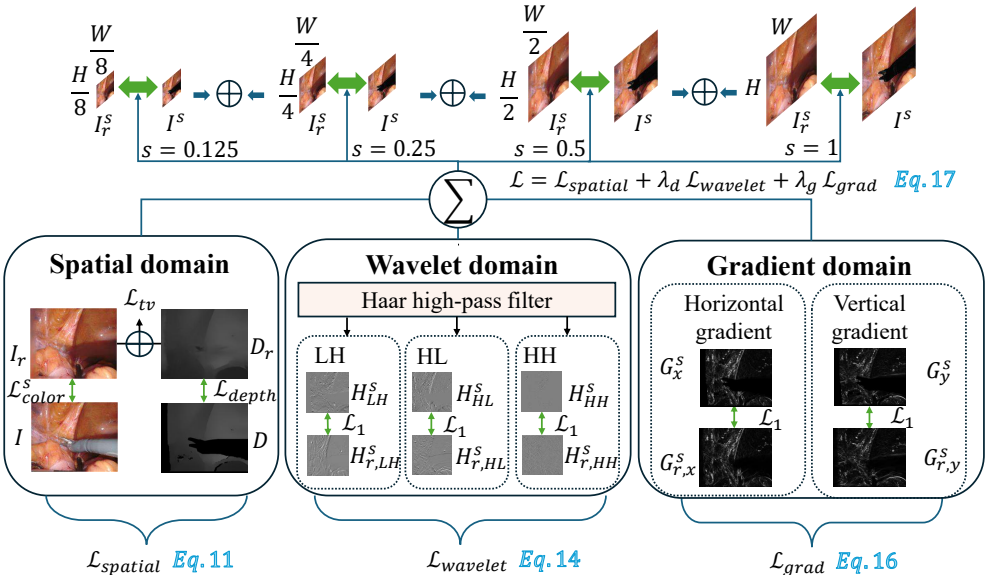


Fig. 4 Mechanism of the proposed Tri-Domain Pyramid Regularization for multi-scale Gaussian refinement.

Spatial Domain. Previous work incorporates standard rendering supervision in the spatial domain to refine the Gaussian representation (Liu et al. 2024). To enforce appearance consistency across different spatial resolutions, we extend the conventional single-scale color supervision to a four-level image pyramid and compute a masked color loss at each scale:

$$\mathcal{L}^s_{\text{color}} = \frac{\|\boldsymbol{M}^s \odot (\boldsymbol{I}^s_r - \boldsymbol{I}^s)\|_1}{\max(\|\boldsymbol{M}^s\|_1, 1)}, \qquad s \in \mathcal{S}, \tag{7}$$

where $\mathcal{S} = \{1, 0.5, 0.25, 0.125\}$ is the set of pyramid scales, $\boldsymbol{I}_r^s$ and $\boldsymbol{I}^s$ are the rendered and ground-truth images bilinearly resized to scale s, respectively, and $\boldsymbol{M}^s$ is the validity mask resized to the corresponding spatial resolution and expanded across the color channels. The denominator averages the accumulated color difference over valid elements, while the maximum operation prevents division by zero when the valid region is empty. The losses obtained at different pyramid levels are combined through a normalized weighted average:

$$\mathcal{L}_{\text{color}} = \frac{\sum_{s\in\mathcal{S}} \omega_s \mathcal{L}_{\text{color}}^s}{\sum_{s\in\mathcal{S}} \omega_s}, \tag{8}$$

where ω_s controls the contribution of scale s. For the four scales in $\mathcal{S}$, the corresponding weights are set to $\{1, 0.5, 0.25, 0.125\}$. Normalizing by the sum of the weights maintains the overall magnitude of the color loss while introducing supervision at progressively coarser spatial resolutions.

Geometric consistency between the rendered and reference depth maps is constrained using a masked correlation loss:

$$\mathcal{L}_{\text{depth}} = 1 - \text{Corr}_{\boldsymbol{M}} \left(\boldsymbol{D}_r, \boldsymbol{D}\right), \tag{9}$$

where $\text{Corr}_{\boldsymbol{M}}(\cdot,\cdot)$ represents the correlation coefficient computed over valid pixels selected by $\boldsymbol{M}$, while $\boldsymbol{D}_r$ and $\boldsymbol{D}$ are the rendered and reference depth maps, respectively. Spatial smoothness is further imposed on the full-resolution rendered color and depth outputs:

$$\mathcal{L}_{\text{tv}} = \text{TV}(\boldsymbol{I}_r) + \text{TV}(\boldsymbol{D}_r). \tag{10}$$

The overall spatial-domain objective combines multi-scale appearance supervision, depth consistency, and spatial smoothness:

$$\mathcal{L}_{\text{spatial}} = \gamma_{\text{color}}\mathcal{L}_{\text{color}} + \gamma_{\text{depth}}\mathcal{L}_{\text{depth}} + \gamma_{\text{tv}}\mathcal{L}_{\text{tv}}, \tag{11}$$

where γ_{color}, γ_{depth}, and γ_{tv} balance the contributions of the color, depth, and total variation terms, respectively.

Wavelet Domain. To suppress high-frequency artifacts while preserving localized anatomical details, we apply a single-level two-dimensional Haar discrete wavelet transform at each level of the shared image pyramid. The validity mask $\boldsymbol{M}^s$ restricts the comparison between the rendered and ground-truth images to valid tissue regions, preventing surgical-tool and invalid pixels from contributing to the wavelet-domain supervision. Since low-frequency appearance and coarse image structure are already constrained by the multi-scale spatial-domain loss, the wavelet-domain objective focuses on discrepancies in the three directional high-frequency components.

For each pyramid scale $s \in \mathcal{S}$, the directional high-frequency coefficient discrepancies are computed by applying the corresponding Haar high-pass kernels to the valid-region rendering residual:

$$\boldsymbol{H}_{r,k}^s - \boldsymbol{H}_k^s = \frac{1}{2}\left[\boldsymbol{M}^s \odot (\boldsymbol{I}_r^s - \boldsymbol{I}^s)\right] *_2 \boldsymbol{K}_k, \qquad k \in \{LH, HL, HH\}, \tag{12}$$

where $\boldsymbol{H}^s_{r,k}$ and $\boldsymbol{H}^s_k$ are the directional high-frequency coefficients associated with the rendered and ground-truth images, respectively, and $\odot$ represents element-wise multiplication, and $*_2$ represents channel-wise filtering with a 2×2 kernel and a stride of two. Owing to the linearity of the Haar transform, this formulation is equivalent to applying the same validity mask to the rendered and ground-truth images before decomposition and subsequently comparing their corresponding high-frequency coefficients. The three Haar high-pass kernels are defined as:

$$\boldsymbol{K}_{LH} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \qquad \boldsymbol{K}_{HL} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \qquad \boldsymbol{K}_{HH} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{13}$$

Under the adopted Haar sign convention, the LH component captures left-to-right intensity variations and is sensitive to vertically oriented boundaries, whereas the HL component captures top-to-bottom intensity variations and responds primarily to horizontally oriented boundaries. The HH component captures diagonal variations associated with corners, fine textures, and localized high-frequency artifacts.

The wavelet-domain loss aggregates the three directional high-frequency coefficient discrepancies across all pyramid scales:

$$\begin{aligned} \mathcal{L}_{\text{wavelet}} = \sum_{s\in\mathcal{S}} \omega_s \Big(& \left\| \boldsymbol{H}^s_{r,LH} - \boldsymbol{H}^s_{LH} \right\|_1 + \left\| \boldsymbol{H}^s_{r,HL} - \boldsymbol{H}^s_{HL} \right\|_1 \\ & + \left\| \boldsymbol{H}^s_{r,HH} - \boldsymbol{H}^s_{HH} \right\|_1 \Big). \end{aligned} \tag{14}$$

The ℓ_1 norm measures the magnitude of the valid-region discrepancy in each directional high-frequency component. Minimizing this loss suppresses localized reconstruction artifacts while preserving anatomical boundaries, tissue textures, and other fine-scale structures.

Gradient Domain. To further preserve anatomical boundaries and tissue contours, we introduce a multi-scale gradient-domain constraint (Liu et al. 2026). At each pyramid scale s, the gradient responses along the horizontal and vertical directions are extracted from the masked rendered and ground-truth images using the corresponding Sobel filters:

$$\boldsymbol{G}^s_{r,d} - \boldsymbol{G}^s_d = \boldsymbol{S}_d * \left[\boldsymbol{M}^s \odot (\boldsymbol{I}^s_r - \boldsymbol{I}^s) \right], \qquad d \in \{x, y\}, \tag{15}$$

where $\boldsymbol{G}^s_{r,d}$ and $\boldsymbol{G}^s_d$ are the directional gradient responses associated with the rendered and ground-truth images, respectively, $\boldsymbol{S}_d$ is the corresponding Sobel filter, and $*$ represents channel-wise convolution. Owing to the linearity of convolution, this formulation is equivalent to applying the same validity mask to the rendered and ground-truth images before gradient extraction and subsequently comparing their corresponding gradient responses. The gradient-domain loss is defined as:

$$\mathcal{L}_{\text{grad}} = \sum_{s\in\mathcal{S}} \omega_s \left(\left\| \boldsymbol{G}^s_{r,x} - \boldsymbol{G}^s_x \right\|_1 + \left\| \boldsymbol{G}^s_{r,y} - \boldsymbol{G}^s_y \right\|_1 \right). \tag{16}$$

The loss compares the horizontal and vertical gradient responses of the rendered and ground-truth images, thereby preserving the strength and orientation of anatomical boundaries. Aggregating these discrepancies across multiple pyramid scales constrains both fine local edges and broader tissue contours, while the ℓ_1 norm reduces sensitivity to isolated gradient outliers.

In summary, the tri-domain pyramid regularization is represented as:

$$\mathcal{L} = \mathcal{L}_{\text{spatial}} + \lambda_d \mathcal{L}_{\text{wavelet}} + \lambda_g \mathcal{L}_{\text{grad}}, \tag{17}$$

where λ_d and λ_g are hyperparameters that balance the contribution of the wavelet-domain and gradient-domain pyramid constraints, respectively. Designed to refine the Gaussian representation beyond conventional spatial supervision, Tri-Domain Pyramid Regularization combines spatial-domain constraints on global appearance and geometry with multi-scale wavelet-domain penalties on localized high-frequency discrepancies and gradient-domain penalties on anatomical boundary misalignment; together, these complementary constraints suppress reconstruction artifacts while preserving tissue contours and fine structural details, thereby enabling more accurate Gaussian refinement and higher-fidelity endoscopic rendering.

4 Experiment

4.1 Experimental Setup

Dataset and Evaluation Metrics. Experiments are conducted on two public endoscopic datasets: SCARED (Allan et al. 2021) and EndoNeRF (Wang et al. 2022). SCARED provides stereo endoscopic sequences of porcine abdominal anatomy captured with the da Vinci system. Following the AF-SfMLearner protocol (Shao et al. 2021), each scene is split into training and testing subsets, and only left-view images are used under the monocular setting. EndoNeRF contains monocular endoscopic video sequences with non-rigid tissue deformation, providing a complementary benchmark for evaluating reconstruction robustness in dynamic surgical scenes. All frames are resized to a fixed resolution for training and evaluation. Reconstruction quality is evaluated using PSNR, SSIM, and LPIPS, where higher PSNR and SSIM values and lower LPIPS values indicate better performance. Metrics are computed for each test frame and then averaged over each sequence.

Implementation Details. For Collaborative Expert Depth Fusion, the Huber threshold is set to $\delta = 1.0$, and the fusion process is performed for $T = 5$ iterations. For Tri-Domain Pyramid Regularization, a four-level image pyramid is constructed using scale factors of $\{1.0, 0.5, 0.25, 0.125\}$. In the wavelet domain, the Haar discrete wavelet transform is applied to penalize discrepancies between corresponding high-frequency sub-bands. In the gradient domain, horizontal and vertical Sobel kernels are applied independently to each channel to compute image gradients at every pyramid level. The wavelet- and gradient-domain loss weights are set to $\lambda_d = 0.2$ and $\lambda_g = 0.35$, respectively. The framework is implemented in PyTorch (Paszke et al. 2017), and its parameters are optimized using Adam. All experiments were conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory.

Table 1 Performance comparison on the EndoNeRF and SCARED datasets.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	Train.Time ↓	GPU ↓
EndoNeRF	EndoNeRF (Wang et al. 2022)	36.062	0.953	0.089	5.0 hours	19 GB
	EndoSurf (Zha et al. 2023)	36.529	0.954	0.074	8.5 hours	17 GB
	EndoGaussian (Liu et al. 2024)	37.849	0.963	0.054	2.0 min	2 GB
	Endo-GSMT (Gou et al. 2025)	38.783	0.968	0.028	– min	– GB
	T2GS (Xu et al. 2025)	32.950	0.921	0.041	3.5 min	6 GB
	Ours	**40.497**	**0.976**	**0.018**	8 min	5 GB
SCARED	EndoNeRF (Wang et al. 2022)	24.345	0.768	0.313	3.5 hours	22 GB
	EndoSurf (Zha et al. 2023)	25.020	0.802	0.356	5.8 hours	22 GB
	EndoGaussian (Liu et al. 2024)	23.477	0.744	0.489	5.01 min	3 GB
	EndoLRMGS (Wang et al. 2025)	27.561	0.861	0.323	– min	– GB
	SG-3DGS (Zhang et al. 2026)	23.550	**0.963**	0.218	– min	– GB
	Ours	**29.719**	0.865	**0.089**	7.0 min	5 GB

4.2 Main Results

Table 1 presents the quantitative comparison on the EndoNeRF and SCARED datasets. Overall, the proposed method achieves the best PSNR and LPIPS on both datasets and the best SSIM on EndoNeRF, while remaining competitive in SSIM on SCARED, demonstrating the effectiveness of combining Collaborative Expert Depth Fusion with Tri-Domain Pyramid Regularization. On the EndoNeRF dataset, our method achieves PSNR, SSIM, and LPIPS scores of 40.497, 0.976, and 0.018, respectively. It outperforms the strongest competing method, Endo-GSMT, by 1.714 dB in PSNR and reduces LPIPS from 0.028 to 0.018. Compared with EndoGaussian, our framework improves PSNR from 37.849 to 40.497 and reduces LPIPS from 0.054 to 0.018. These results demonstrate that Collaborative Expert Depth Fusion provides a geometrically consistent depth prior for Gaussian initialization, while Tri-Domain Pyramid Regularization further enhances photometric fidelity and perceptual consistency during refinement.

On the SCARED dataset, the proposed method obtains a PSNR of 29.719 dB, 0.865 SSIM, and 0.089 LPIPS. Compared with EndoLRMGS, our method improves PSNR by 2.158 and reduces LPIPS from 0.323 to 0.089, showing a clear improvement in perceptual reconstruction quality. Although SG-3DGS reports a higher SSIM, our method achieves substantially better PSNR and LPIPS, suggesting more accurate intensity reconstruction and fewer perceptual artifacts. These results are particularly important for monocular endoscopic reconstruction, where weak textures and specular highlights can easily introduce unstable geometry and local rendering artifacts.

In terms of efficiency, our method remains substantially more efficient than NeRF-based methods, which require several hours of optimization and large GPU

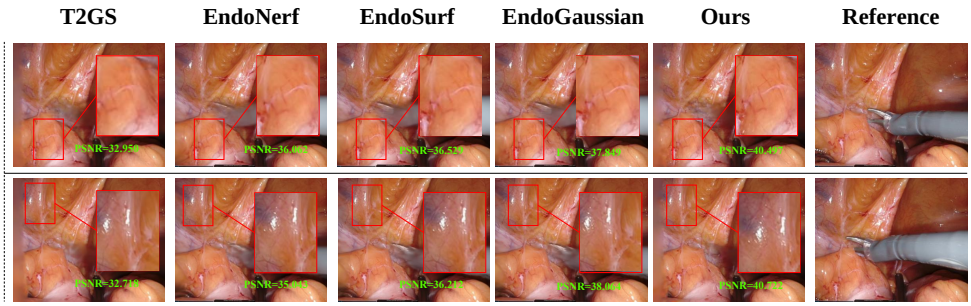


Fig. 5 Qualitative comparison on the EndoNeRF dataset with PSNR annotations. Two representative frames are shown with zoomed-in regions for local detail comparison. The proposed method preserves clearer anatomical structures and achieves higher rendering fidelity than the compared methods.

memory. Compared with lightweight Gaussian-based baselines, the proposed framework introduces moderate additional training cost due to multi-expert depth estimation and tri-domain regularization. However, this overhead is accompanied by clear improvements in reconstruction fidelity and perceptual quality, indicating a favorable quality–efficiency trade-off for robust monocular endoscopic reconstruction.

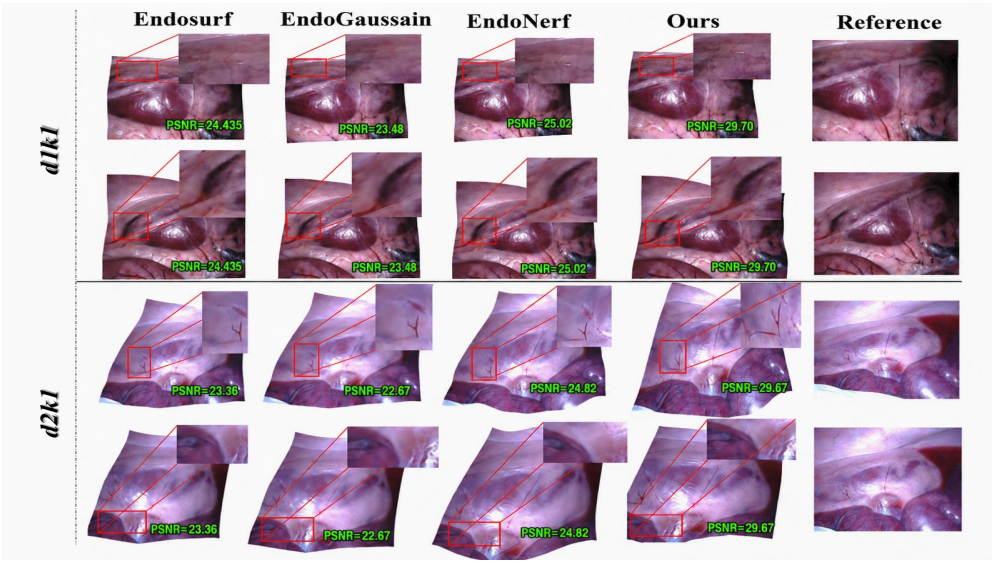


Fig. 6 Qualitative comparison on the SCARED dataset. The proposed method produces clearer structural boundaries and reduces high-frequency artifacts compared with prior methods. The labels d1k1 and d2k1 denote dataset1/keyframe1 and dataset2/keyframe1 in the SCARED dataset, respectively.

The qualitative comparisons in Figs. 5 and 6 are consistent with the quantitative results. Existing methods tend to produce blurred boundaries, degraded local structures, and visible artifacts in weakly textured or highly reflective regions. In contrast, the proposed method reconstructs anatomical contours and local tissue details with fewer visible distortions and high-frequency artifacts. The enlarged regions indicate that the depth prior produced by Collaborative Expert Depth Fusion provides dense initial geometry, while Tri-Domain Pyramid Regularization constrains boundary, contour, and high-frequency discrepancies during Gaussian refinement.

4.3 Depth Prior Evaluation

To directly assess the geometric prior used for Gaussian initialization, we compare the fused depth generated by Collaborative Expert Depth Fusion with representative monocular endoscopic depth estimation methods on the SCARED dataset. All methods are evaluated using the same test split and evaluation protocol. The quantitative results are reported in Table 2.

Table 2 Quantitative comparison of depth-prior quality on the SCARED dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑
DARES (Sheikh Zeinoddin et al. 2024)	0.052	0.356	4.483	0.073	0.980
EndoDAC (Cui et al. 2024)	0.052	0.362	<u>4.464</u>	0.073	0.979
EndoFAST3r (Sheikh Zeinoddin et al. 2025)	<u>0.051</u>	<u>0.354</u>	4.480	–	**0.998**
Ours	**0.050**	**0.353**	**4.437**	**0.071**	<u>0.981</u>

As shown in Table 2, the proposed fused depth prior achieves the lowest Abs Rel, Sq Rel, and RMSE among the compared methods. Among the methods reporting RMSE log, our method also obtains the lowest error. Although EndoFAST3r achieves the highest δ_1 accuracy, the proposed method provides more consistent performance across the error-based metrics. These results indicate that scale-shift alignment and robust collaborative expert fusion reduce expert-specific depth errors and provide a more accurate geometric prior for subsequent Gaussian initialization.

4.4 Ablation Study

Tables 3 and 4 evaluate the effectiveness of the proposed components on the SCARED dataset. The experiments examine two aspects: whether Collaborative Expert Depth Fusion provides more effective geometric guidance than initialization based on a single depth expert, and whether Tri-Domain Pyramid Regularization improves Gaussian refinement beyond single-scale wavelet-domain and gradient-domain constraints.

As shown in Table 3, the baseline using a single-expert depth estimate and the original spatial-domain objective achieves the lowest reconstruction performance.

Table 3 Component-wise ablation study on the SCARED dataset. CEDF denotes Collaborative Expert Depth Fusion, while TDPR denotes Tri-Domain Pyramid Regularization. All configurations use the same depth-guided back-projection procedure for Gaussian initialization.

Components	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline	26.092	0.696	0.422
CEDF	28.437	0.797	0.300
CEDF + Wavelet Reg.	29.195	0.832	0.210
CEDF + Wavelet Reg. + Gradient Reg.	29.700	0.850	0.126
CEDF + TDPR	**29.719**	**0.865**	**0.089**

Introducing Collaborative Expert Depth Fusion substantially improves all three metrics, indicating that aligning and integrating complementary expert depth estimates provides a more effective geometric basis for Gaussian initialization while reducing the influence of locally inconsistent predictions. Adding single-scale wavelet-domain regularization further improves reconstruction fidelity and perceptual quality by constraining high-frequency discrepancies. Incorporating the gradient-domain constraint produces additional gains by preserving anatomical boundaries and tissue contours. The complete model with Tri-Domain Pyramid Regularization achieves the best overall performance, demonstrating that applying complementary constraints across multiple image scales is more effective than single-scale wavelet-domain and gradient-domain regularization for Gaussian refinement.

Table 4 Ablation study of different expert combinations in Collaborative Expert Depth Fusion. Experts A, B, and C represent three complementary monocular depth estimators whose predictions are aligned through scale-shift fitting and integrated through residual-dependent weighting.

Components	PSNR ↑	SSIM ↑	LPIPS ↓
Expert A	29.676	0.861	0.116
Expert A + Expert B	29.698	0.863	0.105
Expert A + Expert C	29.684	0.859	0.112
Expert A + Expert B & C	**29.719**	**0.865**	**0.089**

Table 4 further evaluates the complementarity of different depth experts under the same Gaussian refinement setting. Expert A alone provides a strong result, while incorporating either Expert B or Expert C yields further improvements, indicating that the additional experts contribute complementary geometric information. The different gains obtained by combining Expert A with Expert B or Expert C suggest that the depth estimates produced by the two experts are not redundant and capture different structural characteristics. Combining all three experts achieves the best overall performance, confirming that Collaborative Expert Depth Fusion benefits from integrating complementary depth estimates while suppressing locally inconsistent predictions.

4.5 Hyperparameter Analysis

The proposed Tri-Domain Pyramid Regularization introduces two weighting factors, λ_d and λ_g, to balance the wavelet-domain and gradient-domain constraints in Eq. (17). To investigate their impact on reconstruction performance, we conduct a sensitivity analysis on the SCARED dataset by varying one weight while keeping the other settings fixed. The quantitative results are presented in Fig. 7.

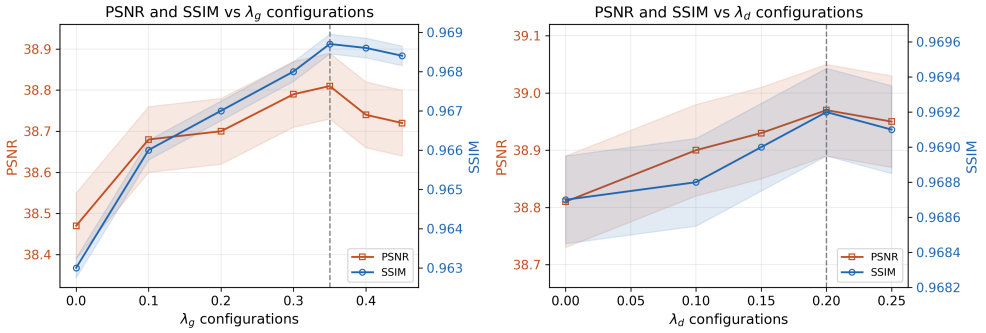


Fig. 7 Impact of hyperparameters on reconstruction performance. PSNR and SSIM variations with respect to λ_g (left) and λ_d (right). Both metrics first improve and then slightly decrease as the regularization weight increases, indicating that moderate regularization achieves the best performance among the tested settings. The vertical dashed lines denote the selected default values.

Wavelet-Domain Weight λ_d. The parameter λ_d controls the strength of the wavelet-domain regularization. When λ_d is too small, the constraint provides insufficient supervision over localized high-frequency responses, limiting its ability to suppress artifacts. As λ_d increases to a moderate range, both reconstruction fidelity and perceptual quality improve, indicating that frequency-domain regularization helps stabilize fine-detail reconstruction. However, an excessively large λ_d may over-penalize high-frequency variations and suppress useful anatomical details. As shown in Fig. 7, the best empirical performance among the tested settings is achieved at $\lambda_d = 0.2$, which is adopted as the default value.

Gradient-Domain Weight λ_g. The coefficient λ_g controls the contribution of the gradient-domain loss to the overall optimization objective. When λ_g is small, the gradient-domain term contributes weakly to the overall objective, resulting in limited constraints on anatomical boundaries and tissue contours. Increasing λ_g places greater emphasis on discrepancies along anatomical boundaries and tissue contours, which can produce sharper reconstructed boundaries. However, an excessively large λ_g may cause the gradient-domain term to dominate the optimization objective, reducing photometric fidelity and introducing local structural distortions. As shown in Fig. 7, $\lambda_g = 0.35$ achieves the highest overall performance among the evaluated settings and is therefore used in all experiments.

Overall, both weighting factors exhibit a consistent trend: moderate regularization improves reconstruction quality, whereas overly weak or overly strong constraints

degrade performance. This analysis confirms that the wavelet-domain and gradient-domain constraints are most effective when properly balanced with the spatial-domain reconstruction objective.

5 Discussion

5.1 Analysis of the Proposed Components

The results in Tables 2–4 show that Collaborative Expert Depth Fusion and Tri-Domain Pyramid Regularization improve complementary stages of Gaussian reconstruction. Collaborative Expert Depth Fusion reduces several depth errors and outperforms single-expert initialization by mapping complementary predictions into a common scale-shift space and suppressing locally inconsistent estimates through residual-dependent weighting. The resulting fused depth serves as relative geometric guidance rather than an estimate of absolute metric depth. Its larger improvement in LPIPS indicates that the fusion primarily enhances local structural and perceptual consistency.

Tri-Domain Pyramid Regularization further improves the initialized representation by imposing complementary constraints across spatial, wavelet, and gradient domains. Spatial-domain supervision preserves global appearance and geometry, while wavelet- and gradient-domain constraints reduce high-frequency discrepancies and preserve anatomical boundaries. The ablation and hyperparameter results also show that these constraints must be appropriately balanced, as excessive weights may suppress valid local variations or overemphasize edge agreement.

5.2 Practical Considerations and Limitations

The proposed framework requires approximately 7-8 minutes and 5 GB of GPU memory, remaining substantially more efficient than the evaluated NeRF-based methods but introducing additional cost compared with lightweight Gaussian baselines. This overhead arises mainly from multi-expert depth inference and multi-scale regularization, making the current implementation more suitable for per-scene reconstruction than continuous intraoperative updating.

The evaluation is limited to SCARED and EndoNeRF and primarily relies on image-space reconstruction and depth metrics, which do not directly measure surface accuracy, topology, or temporal geometric consistency. The framework also uses fixed depth experts, fusion weights, pyramid scales, and calibrated camera parameters. Future work will investigate uncertainty-aware expert fusion, adaptive domain weighting, joint camera and geometry refinement, and evaluation on more diverse sequences with stronger deformation, occlusion, and topology changes.

6 Conclusion

This work presents a robust depth-prior-guided Gaussian reconstruction framework for monocular endoscopic scenes. Collaborative Expert Depth Fusion improves Gaussian initialization by aligning predictions from multiple monocular depth experts

through scale-shift fitting and robustly integrating the aligned predictions to construct a fused depth prior. This prior is subsequently back-projected to provide dense geometric support for Gaussian construction. To further improve refinement quality, Tri-Domain Pyramid Regularization introduces complementary constraints in the spatial, wavelet, and gradient domains, enabling the optimization to preserve global appearance, suppress localized high-frequency artifacts, and maintain anatomical boundary consistency. Experiments on the SCARED and EndoNeRF datasets demonstrate that the proposed framework consistently improves reconstruction fidelity and perceptual quality over existing endoscopic reconstruction methods. Ablation studies further validate the effectiveness of Collaborative Expert Depth Fusion and Tri-Domain Pyramid Regularization, showing that geometrically consistent initialization and multi-domain refinement are both important for robust monocular endoscopic Gaussian reconstruction. Future work will explore stronger depth-prior construction and extend the framework to more complex dynamic surgical scenarios.

References

Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., Zhu, Z., Luo, H., Jia, F., Zhang, X., Li, X., Sharan, L., Kurmann, T., Schmid, S., Sznitman, R., Psychogyios, D., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S.: Stereo Correspondence and Reconstruction of Endoscopic Data Challenge (2021). https://arxiv.org/abs/2101.01133

Bhat, S.F., Alhashim, I., Wonka, P.: LocalBins: Improving Depth Estimation by Learning Local Distributions. In: European Conference on Computer Vision, pp. 480–496. Springer, Cham (2022)

Bidokh, E., Hassanpour, H.: Enhancing Wireless Capsule Endoscopy Images from Intense Illumination Specular Reflections Using The Homomorphic Filter. Biomedical Signal Processing and Control **84**, 104723 (2023)

Brandao, P., Psychogyios, D., Mazomenos, E., Stoyanov, D., Janatka, M.: HAP-Net: Hierarchically Aggregated Pyramid Network for Real-Time Stereo Matching. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **9**(3), 219–224 (2021)

Cui, B., Islam, M., Bai, L., Wang, A., Ren, H.: Endodac: Efficient Adapting Foundation Model for Self-Supervised Depth Estimation from Any Endoscopic Camera. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 208–218. Springer, Cham (2024)

Chen, L., Tang, W., John, N.W., Wan, T.R., Zhang, J.J.: SLAM-Based Dense Surface Reconstruction in Monocular Minimally Invasive Surgery and Its Application to Augmented Reality. Computer methods and programs in biomedicine **158**, 135–146 (2018)

Chong, N., Yang, F.: A Monocular Medical Endoscopic Images Depth Estimation

Method Based on A Confidence-Guided Dual-Branch Siamese Network. Biomedical Signal Processing and Control **102**, 107–123 (2025)

Grasa, O.G., Bernal, E., Casado, S., Gil, I., Montiel, J.: Visual SLAM for Handheld Monocular Endoscope. IEEE transactions on medical imaging **33**(1), 135–146 (2013)

Gou, H., Wang, C., Yang, J., Liu, Y., Jia, F., Xiao, D., Qin, F., Luo, H.: Endo-GSMT: Endoscopic Monocular Scene Reconstruction with Dynamic Gaussian Splatting and Motion Tracking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 213–223. Springer, Cham (2025)

Huang, Y., Bai, L., Cui, B., Li, Y., Chen, T., Wang, J., Wu, J., Lei, Z., Liu, H., Ren, H.: Endo-4DGX: Robust Endoscopic Scene Reconstruction and Illumination Correction with Gaussian Splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 181–191. Springer, Cham (2025)

Huang, Y., Cui, B., Bai, L., Guo, Z., Xu, M., Islam, M., Ren, H.: Endo-4DGS: Endoscopic Monocular Scene Reconstruction with 4D Gaussian Splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 197–207. Springer, Cham (2024)

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., *et al.*: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)

Liu, X., Li, Z., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Sage: SLAM with Appearance and Geometry Prior for Endoscopy. In: 2022 IEEE International Conference on Robotics and Automation (ICRA), pp. 5587–5593. IEEE, Piscataway (2022)

Liu, Y., Li, C., Yang, C., Yuan, Y.: EndoGaussian: Real-time Gaussian Splatting for Dynamic Endoscopic Scene Reconstruction (2024). https://arxiv.org/abs/2401.12561

Li, Y., Richter, F., Lu, J., Funk, E.K., Orosco, R.K., Zhu, J., Yip, M.C.: Super: A Surgical Perception Framework for Endoscopic Tissue Manipulation with Surgical Robotics. IEEE Robotics and Automation Letters **5**(2), 2294–2301 (2020)

Liu, X., Stiber, M., Huang, J., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Reconstructing Sinus Anatomy from Endoscopic Video–Towards a Radiation-Free Approach for Quantitative Longitudinal Assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 3–13. Springer, Cham (2020)

Liu, J., Zhang, Q., Wan, Q., Cheng, X., Geng, S., Ding, W.: DMNet: Detail Modulation Network with Gradient-Texture Guidance for Lightweight Image Super-Resolution. Journal of King Saud University Computer and Information Sciences **38**(4), 146

(2026)

Marcus, H.J., Cundy, T.P., Hughes-Hallett, A., Yang, G.-Z., Darzi, A., Nandi, D.: Endoscopic and Keyhole Endoscope-Assisted Neurosurgical Approaches: A Qualitative Survey on Technical Challenges and Technological Solutions. British journal of neurosurgery **28**(5), 606–610 (2014)

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. Communications of the ACM **65**(1), 99–106 (2021)

Mahendran, A., Vedaldi, A.: Understanding Deep Image Representations by Inverting Them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5188–5196. IEEE, Piscataway (2015)

Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., *et al.*: EndoSLAM Dataset and An Unsupervised Monocular Visual Odometry and Depth Estimation Approach for Endoscopic Videos. Medical image analysis **71**(1), 102058 (2021)

Peng, J., Chen, C., Liu, Y.: F2Plenoxels: Fast Voxel Radiance Fields without Neural Networks for Few-Shot View Synthesis. Journal of King Saud University Computer and Information Sciences **38**(3), 98 (2026)

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic Differentiation in PyTorch. In: Neural Information Processing Systems Workshop, Long Beach (2017)

Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 519–528. IEEE, Piscataway (2006)

Schonberger, J.L., Frahm, J.-M.: Structure-from-Motion Revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113. IEEE, Piscataway (2016)

Shao, S., Pei, Z., Chen, W., Zhang, B., Wu, X., Sun, D., Doermann, D.: Self-Supervised Learning for Monocular Depth Estimation on Minimally Invasive Surgery Scenes. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 7159–7165. IEEE, Piscataway (2021)

Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B.: Self-Supervised Monocular Depth and Ego-Motion Estimation in Endoscopy: Appearance Flow to the Rescue. Medical image analysis **77**, 102338 (2022)

Stoyanov, D.: Surgical Vision. Annals of biomedical engineering **40**(2), 332–345 (2012)

Song, X., Tang, H., Yang, C., Zhou, G., Wang, Y., Huang, X., Hua, J., Coatrieux, G., He, X., Chen, Y.: Deformable Transformer for Endoscopic Video Super-Resolution. Biomedical Signal Processing and Control **77**, 103827 (2022)

Sheikh Zeinoddin, M., Hoque, M.I., Tandogdu, Z., Shaw, G.L., Clarkson, M.J., Mazomenos, E.B., Stoyanov, D.: Endo-Fast3R: Endoscopic Foundation Model Adaptation for Structure from Motion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 117–126. Springer, Cham (2025)

Sheikh Zeinoddin, M., Lena, C., Qu, J., Carlini, L., Magro, M., Kim, S., De Momi, E., Bano, S., Grech-Sollars, M., Mazomenos, E., *et al.*: DARES: Depth Anything in Robotic Endoscopic Surgery with Self-Supervised Vector-LoRA of the Foundation Model. In: European Conference on Computer Vision, pp. 1–11. Springer, Cham (2024)

Shao, S., Zhu, K., Ma, S., Du, X., Zhang, B., Min, Z.: Depth Any Endoscopy: Towards Self-Supervised Generalizable Depth Estimation in Monocular Endoscopy. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 34126–34137. IEEE, Piscataway (2026)

Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural Rendering for Stereo 3D Reconstruction of Deformable Tissues in Robotic Surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 431–441. Springer, Cham (2022)

Wang, X., Zhang, S., Huang, B., Stoyanov, D., Mazomenos, E.B.: EndoLRMGS: Complete Endoscopic Scene Reconstruction Combining Large Reconstruction Modelling and Gaussian Splatting (2025). https://arxiv.org/abs/2503.22437

Xu, J., Li, C., Liu, P., Pfeiffer, M., Liu, L., Docea, R., Wagner, M., Speidel, S.: T 2 GS: Comprehensive Reconstruction of Dynamic Surgical Scenes with Gaussian Splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 595–605. Springer, Cham (2025)

Xiang, Y., Xiao, W., Li, W., Chang, Q., Liu, K.: MSLP-S2MDEL: Multi-Scale Laplacian Pyramid-Based Self-Supervised Monocular Depth Estimation for Laparoscopic Surgery. Biomedical Signal Processing and Control **119**, 109890 (2026)

Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 20331–20341. IEEE, Piscataway (2024)

Yang, Z., Pan, J., Dai, J., Sun, Z., Xiao, Y.: Self-Supervised Endoscopy Depth Estimation Framework with CLIP-Guidance Segmentation. Biomedical Signal Processing and Control **95**, 106410 (2024)

Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: EndoSurf: Neural Surface Reconstruction of Deformable Tissues with Stereo Endoscope Videos. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 13–23. Springer, Cham (2023)

Zhou, Z., Fan, X., Shi, P., Xin, Y., Zhang, C., Zhang, Y.: R-TAFM: Purely Convolutional Recurrent Refinement for Deployment-Oriented Monocular Depth Estimation. Journal of King Saud University Computer and Information Sciences **38**(3), 108 (2026)

Zhou, H., Jagadeesan, J.: Real-Time Dense Reconstruction of Tissue Surface from Stereo Optical Video. IEEE transactions on medical imaging **39**(2), 400–412 (2019)

Zhou, H., Jayender, J.: EMDQ-SLAM: Real-Time High-Resolution Reconstruction of Soft Tissue Surface from Stereo Laparoscopy Videos. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 331–340. Springer, Cham (2021)

Zhong, J., Ren, H., Chen, Q., Zhang, H.: A Review of Deep Learning-Based Localization, Mapping and 3D Reconstruction for Endoscopy. Journal of Micro and Bio Robotics **21**(1), 1 (2025)

Zhang, Z., Song, H., Fan, J., Shao, L., Fu, T., Ai, D., Xiao, D., Wang, Y., Lin, Y., Yang, J.: SG-3DGS: Sequential Growing 3D Gaussian Splatting for Scene Reconstruction of Monocular Endoscope Video. IEEE Transactions on Medical Imaging **45**(5), 1802–1816 (2026)

Zhu, K., Shao, S., Yang, Y., Tian, Z., Zhang, B., Min, Z.: Beyond Foundation Models: Distilling Geometric Priors for Lightweight Monocular Depth Estimation in Endoscopy. IEEE Transactions on Medical Imaging **45**(7), 4002–4013 (2026)

Zhu, L., Wang, Z., Cui, J., Jin, Z., Lin, G., Yu, L.: EndoGS: Deformable Endoscopic Tissues Reconstruction with Gaussian Splatting. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 135–145. Springer, Cham (2024)

Statements and Declarations

Consent for publication

Not applicable.

Data availability

The data used in this study is publicly available from the SCARED and EndoNeRF datasets.

Materials availability

Not applicable.

Code availability

The code will be made available upon publication.