Rendering with a Gut Feeling: Depth-Guided Triangle Splatting for Physically Consistent Colonoscopic Reconstruction

Romain Hardy[1] ⬤   ROMAIN_HARDY@FAS.HARVARD.EDU

Andrea Dunn Beltran[2]   ASDUNNBE@CS.UNC.EDU

Todd A. Brenner[3]

Tyler M. Berzin[3]

Pranav Rajpurkar[1]

[1] *Harvard University, Department of Biomedical Informatics, Boston, MA, USA*

[2] *The University of North Carolina, Department of Computer Science, Chapel Hill, NC, USA*

[3] *Beth Israel Deaconess Medical Center, Boston, MA, USA*

Abstract

Colonoscopy scene reconstruction under monocular imaging remains challenging due to affine depth ambiguity in geometric priors and strong viewpoint-dependent specularities from coaxial illumination. We present **GutSee**, a depth-guided triangle splatting framework that addresses these challenges through two key innovations. First, we introduce an *affine-invariant* depth supervision scheme that accounts for per-frame scale and shift ambiguities in pretrained monocular depth estimators, enabling them to provide stable geometric guidance even when their predictions are mutually inconsistent. Second, we incorporate a physically motivated illumination model with an explicit coaxial spotlight and learnable BRDF parameters, preventing specular highlights from being misinterpreted as geometry. Together with triangle primitives that naturally enforce surface continuity, these components yield reconstructions that are both geometrically faithful and photometrically realistic. On a phantom colonoscopy dataset, GutSee reduces mean depth RMSE by 18.3% over the next-best method under biased supervision while maintaining comparable rendering quality. These results demonstrate that coupling affine-invariant depth guidance with physically accurate lighting models improves resilience to supervision bias, enabling reliable reconstruction even when using imperfect depth priors.

Keywords: Endoscopy, 3D reconstruction, neural rendering.

1. Introduction

Colonoscopy is the clinical gold standard for colorectal cancer screening, yet endoscopists must infer complex anatomy from monocular video alone (Rex et al., 2017). In the absence of explicit three-dimensional (3D) cues, perceptual judgments remain qualitative and error-prone, contributing to incomplete mucosal coverage, imprecise lesion localization, and unreliable size estimation (Van Rijn et al., 2006; O'Connor et al., 2016; Schoen et al., 1997; Anderson et al., 2016). Robust 3D reconstruction offers a means to mitigate these limitations by providing navigational context (Wang et al., 2024b), supporting quantitative assessment (Abdelrahim et al., 2022; Soriero et al., 2022), and enabling comprehensive surface mapping (Ma et al., 2019; Thompson et al., 2016). However, the colonic environment differs substantially from conventional computer vision settings (Zhou et al., 2024).

Monocular reconstruction in colonoscopy is challenging for several reasons. Scale is inherently ambiguous without additional constraints (Wang et al., 2023). Tissue deformation arising from insufflation and peristalsis disrupts geometric consistency, and mucosal textures are often repetitive or locally feature-poor, complicating correspondence-based approaches (Ma et al., 2021; Schmidt et al., 2024). Long exploratory trajectories further exacerbate drift, yielding accumulated inconsistencies over the course of an examination (Pyykölä et al., 2024; Yao et al., 2021; Golhar et al., 2025). Illumination adds yet another layer of difficulty: the coaxial light source generates strong, pose-dependent specular reflections that violate the Lambertian assumptions underlying most photometric objectives (Shandro et al., 2020).

Figure 1: GutSee substantially improves geometric reconstruction accuracy under biased priors while preserving strong photometric fidelity.

Differentiable scene representations such as Neural Radiance Fields (Mildenhall et al., 2021) and Gaussian Splatting (Kerbl et al., 2023) have recently advanced general-purpose 3D reconstruction (Yan et al., 2024). These models recover geometry by minimizing photometric discrepancies between rendered and observed frames, but their flexibility can allow view-dependent appearance and density fields to absorb modeling errors, yielding photorealistic renderings with inaccurate underlying geometry (Huang et al., 2024; Wang et al., 2024a). In clinical endoscopy, where downstream utility depends on accurate surface reconstruction rather than appearance alone, such behavior is problematic.

A common approach to improving geometric fidelity is to incorporate depth supervision, typically from dedicated sensors or annotated datasets (Wang et al., 2025). In endoscopy, however, acquiring metric depth in vivo is infeasible, and even phantom datasets require careful calibration. Monocular depth networks pretrained on large-scale data offer an attractive alternative (Yang et al., 2024; Hardy et al., 2025), as they provide geometrically coherent *relative* depth. Their predictions, though, are only defined up to an affine transformation. Directly supervising rendered depth with such signals can therefore introduce undesirable bias or instability, motivating the need for a principled mechanism to exploit affine-ambiguous priors without imposing incorrect scale or shift.

A second obstacle is the entanglement of lighting and geometry in photometric losses. Standard diffuse reflectance models cannot explain the prominent, viewpoint-dependent specularities induced by coaxial endoscope lighting (Psychogyios et al., 2023; Batlle et al., 2023; Han and Wu, 2025). As a result, reconstruction methods may inadvertently encode specular highlights as geometric features unless illumination is modeled explicitly.

We introduce **GutSee**, a depth-guided triangle splatting framework designed to overcome these challenges in colonoscopic reconstruction. By combining differentiable triangle primitives with a physically based illumination model, GutSee recovers coherent surface ge-
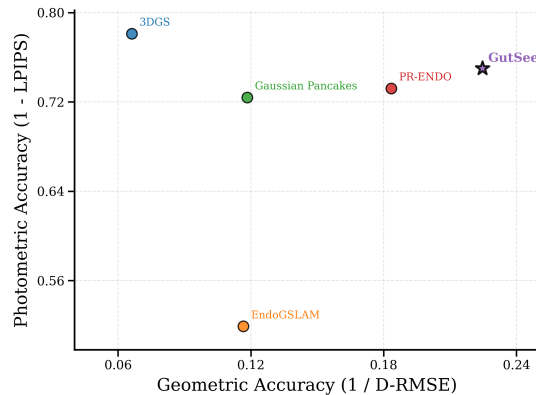
ometry and realistic appearance from monocular video and affine-ambiguous depth priors. Our contributions are:

- An affine-invariant supervision scheme for rendered depth and normals that enables effective geometric guidance from pretrained monocular depth estimators, tolerating residual frame-wise scale and shift inconsistencies in the prior.

- A physically motivated lighting model based on a bidirectional reflectance distribution function (BRDF) with learnable material parameters and an explicit coaxial spotlight, which accounts for endoscope-specific illumination and prevents specular reflections from introducing spurious geometric detail.

- State-of-the-art geometric reconstruction accuracy on a phantom colonoscopy benchmark among methods supervised with biased depth priors, reducing the gap to methods supervised with ground-truth metric depth while maintaining high-fidelity novel view synthesis.

2. Related Work

Explicit Differentiable Representations. Recent neural rendering methods increasingly rely on explicit primitives that can be rasterized efficiently while remaining differentiable. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) exemplifies this trend, representing scenes as anisotropic Gaussians jointly optimized for appearance and geometry. While effective for view synthesis, the infinite support of Gaussians can blur fine structures and introduce view-inconsistent surfaces, motivating the use of primitives with tighter spatial support—such as 2D Gaussians (Huang et al., 2024), smooth convex shapes (Held et al., 2025b), and Triangle Splatting (Held et al., 2025a), which offers mesh-compatible geometry with well-defined normals. We build on Triangle Splatting, whose explicit surface parameterization supports geometric priors and physically consistent illumination modeling.

Depth Supervision Under Ambiguous Geometry. Monocular endoscopy poses fundamental difficulties for geometric recovery, as metric depth is generally unavailable. Existing neural rendering approaches address this by supervising geometry with learned priors (Bonilla et al., 2024; Ma et al., 2021), but the resulting reconstructions inherit the biases of these signals. Real-time systems such as EndoGSLAM (Wang et al., 2024b) mitigate drift through SLAM-style tracking and structured map updates, yet remain vulnerable to errors originating from imperfect depth supervision. Our work introduces affine-invariant depth supervision to reconcile per-frame scale and shift ambiguities, allowing relative depth structure to guide optimization without the instability of metrically constrained but biased losses.

Illumination Modeling for Near-Field Imaging. A second obstacle to reliable geometric recovery is the near-field, coaxial illumination characteristic of endoscopy. Strong specularities and radial attenuation systematically violate the distant-light assumptions used in many photometric objectives, causing renderers to misinterpret lighting artifacts as geometry. Prior efforts incorporate endoscope-specific lighting (Kaleta et al., 2025; Han and Wu, 2025), but they depend on accurate geometry or MLP-based irradiance estimation and
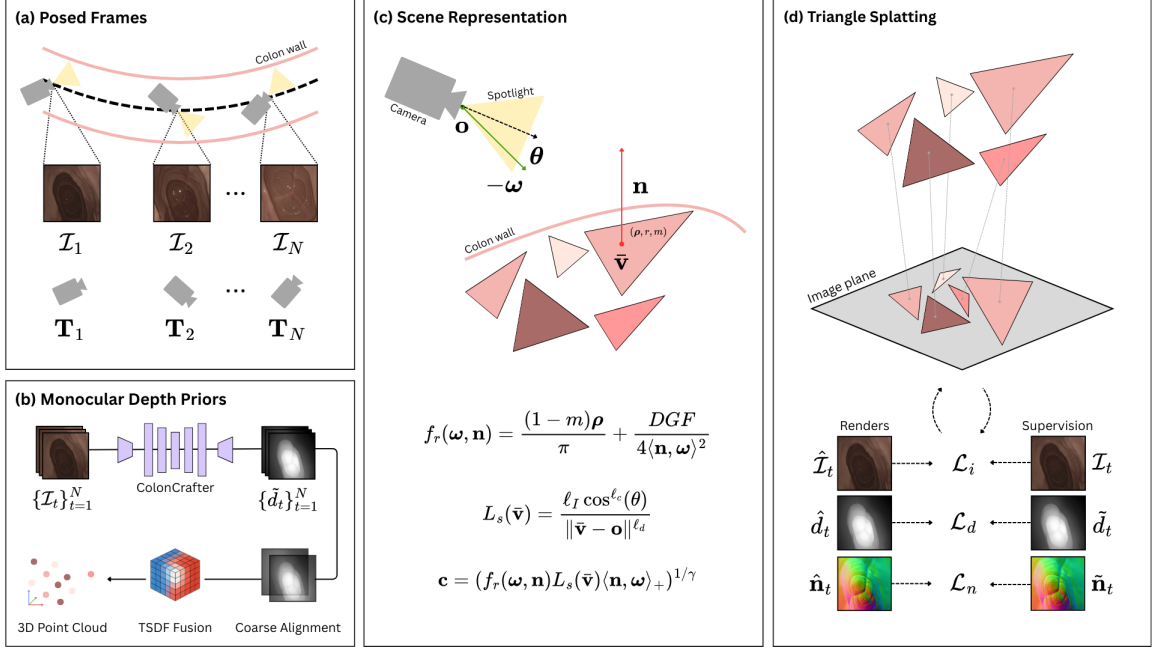
Figure 2: **Overview of GutSee. (a)** Given a sequence of posed RGB colonoscopy frames, **(b)** we estimate disparity maps using ColonCrafter (Hardy et al., 2025) and fuse them into an initial point cloud via TSDF integration. **(c)** The scene is represented as a set of triangle primitives with learnable material properties, rendered under a physically-based illumination model that combines a coaxial spotlight with low-frequency ambient lighting. **(d)** Triangle geometry, appearance, and lighting are optimized using photometric losses and affine-invariant geometric supervision against the biased ColonCrafter priors.

therefore remain sensitive to biased priors. We address this complementary failure mode by integrating a learnable BRDF and explicit coaxial spotlight into our splatting framework, enabling correct attribution of view-dependent effects and improving geometric robustness under imperfect supervision.

3. Methods

Given RGB colonoscopy frames $\{\mathcal{I}_t\}_{t=1}^N$ with corresponding camera poses $\{\mathbf{T}_t\}_{t=1}^N$, we reconstruct the colonic surface by optimizing a set of small triangles to match both the observed images and depth priors from a pretrained monocular estimator. The triangles carry material properties that interact with an explicit model of the endoscope's coaxial light source, allowing specular highlights to be explained by lighting rather than absorbed into geometry. Because monocular depth is only recoverable up to an unknown scale and shift, we supervise geometry using an affine-invariant loss that aligns rendered and predicted depths before computing residuals.

We describe the triangle-based scene representation (Section 3.1), the physically based illumination model (Section 3.2), and the optimization procedure with affine-invariant supervision (Section 3.3). An overview is shown in Figure 2.

3.1. Scene Representation

We represent the colon surface as a collection of small, overlapping triangles rendered via differentiable splatting. Each triangle stores its geometry (vertex positions), transparency, and material properties (color, roughness, shininess). During rendering, triangles are projected onto the image plane with soft edges that enable gradient-based optimization, then composited front-to-back to produce the final image. Because each triangle defines a well-posed local surface patch, this representation naturally provides explicit surface normals, which will be used in our physically based illumination model.

Formally, we modify the Triangle Splatting framework (Held et al., 2025a) and represent the scene as a set of triangle primitives $\{T^{(k)}\}_{k=1}^{K}$. Each triangle $T^{(k)}$ is parameterized by vertices $\{\mathbf{v}_i^{(k)}\}_{i=1}^{3}$, an opacity $\alpha^{(k)}$, a smoothness parameter $\sigma^{(k)}$, and material attributes: diffuse albedo $\boldsymbol{\rho}^{(k)}$, roughness $r^{(k)}$, and metallicity $m^{(k)}$. Rendering proceeds by projecting vertices to image space via a pinhole camera. Each triangle contributes a soft footprint given by a differentiable window function $w^{(k)}(\mathbf{u}; \sigma^{(k)})$ depending on the signed distance from pixel $\mathbf{u}$ to the projected triangle, and final pixel intensities are obtained by front-to-back alpha compositing. The geometric normal

$$\mathbf{n}^{(k)} = \frac{(\mathbf{v}_2^{(k)} - \mathbf{v}_1^{(k)}) \times (\mathbf{v}_3^{(k)} - \mathbf{v}_1^{(k)})}{\|(\mathbf{v}_2^{(k)} - \mathbf{v}_1^{(k)}) \times (\mathbf{v}_3^{(k)} - \mathbf{v}_1^{(k)})\|}$$

of $T^{(k)}$ is flipped to face the camera for view consistency, and shading is evaluated at the triangle centroid $\bar{\mathbf{v}}^{(k)}$.

3.2. Endoscopic Illumination Model

Standard rendering models assume distant or ambient lighting, but endoscopes carry a coaxial light source that moves with the camera, producing characteristic radial falloff and view-dependent specular highlights. Following prior physically based rendering formulations, we treat the endoscopic light as a camera-mounted spotlight and describe surface reflectance using a microfacet BRDF. In particular, we adapt a Disney-style BRDF parameterization (Burley and Studios, 2012), drawing inspiration from Han and Wu (2025) and Yao et al. (2022). By modeling material reflectance and illumination separately, the optimization avoids confusing specular highlights with changes in surface geometry.

The illumination has two components. The spotlight intensity at a surface point falls off with both distance from the camera and angle from the optical axis, controlled by learnable parameters. The surface reflection combines a diffuse term (light scattered equally in all directions, controlled by albedo) and a specular term (mirror-like reflection concentrated around the reflection direction, controlled by roughness and metallicity). Because the light source and camera are co-located, incoming and outgoing light directions coincide, simplifying the reflection computation.

Formally, given camera center $\mathbf{o}$, the viewing direction at centroid $\bar{\mathbf{v}}$ is

$$\boldsymbol{\omega} = \frac{\bar{\mathbf{v}} - \mathbf{o}}{\|\bar{\mathbf{v}} - \mathbf{o}\|}.$$

The spotlight radiance is

$$L_s(\bar{\mathbf{v}}) = \frac{\ell_I \cos^{\ell_c}(\theta)}{\|\bar{\mathbf{v}} - \mathbf{o}\|^{\ell_d}},$$

where θ is the angle to the optical axis and $\ell_I, \ell_c, \ell_d > 0$ are learnable intensity, angular falloff, and distance attenuation exponents. The surface BRDF is

$$f_r(\boldsymbol{\omega}, \mathbf{n}) = \underbrace{\frac{(1-m)\,\boldsymbol{\rho}}{\pi}}_{\text{diffuse}} + \underbrace{\frac{D(r, \mathbf{n}, \boldsymbol{\omega})\,G(r, \mathbf{n}, \boldsymbol{\omega})\,F(\boldsymbol{\omega}, \mathbf{n}, m, \boldsymbol{\rho})}{4\,\langle \mathbf{n}, \boldsymbol{\omega} \rangle^2}}_{\text{specular}},$$

where D is the GGX normal-distribution function, G is the Smith geometry term, and F is the Schlick Fresnel approximation (Yao et al., 2022). We compute the outgoing radiance as

$$\mathbf{c} = \left(f_r(\boldsymbol{\omega}, \mathbf{n}) L_s(\bar{\mathbf{v}}) \langle \mathbf{n}, \boldsymbol{\omega} \rangle_+ \right)^{1/\gamma},$$

where γ is a learnable gamma correction mapping linear radiance to the nonlinear image domain.

3.3. Optimization

We initialize geometry from a point cloud produced by fusing ColonCrafter (Hardy et al., 2025) depth predictions via TSDF integration, then jointly optimize triangle geometry, material properties, and lighting parameters. The loss combines photometric reconstruction with affine-invariant geometric supervision.

The core challenge is that monocular depth estimators predict depth only up to an unknown scale and shift that may vary across frames. Rather than treating these predictions as metric ground truth, we align rendered depths to the predictions before computing residuals: for each frame, we solve for the scale and shift that best match rendered depth to predicted depth, then penalize deviations from the aligned target. This allows the depth prior to guide relative geometry without imposing an incorrect scale. We apply analogous supervision to surface normals derived from the aligned depths, encouraging local surface orientation to remain consistent with the prior.

Formally, the photometric loss combines pixel-wise and structural similarity terms:

$$\mathcal{L}_i = (1 - \lambda_{\text{SSIM}})\|\hat{\mathcal{I}} - \mathcal{I}\|_1 + \lambda_{\text{SSIM}}(1 - \text{SSIM}(\hat{\mathcal{I}}, \mathcal{I})).$$

For geometric supervision, we align each rendered disparity map $\hat{d}(\mathbf{u})$ to the ColonCrafter prior $\tilde{d}(\mathbf{u})$ by solving a per-view least-squares fit:

$$(\beta^*, \xi^*) = \arg\min_{\beta, \xi} \sum_{\mathbf{u} \in \Omega} (\beta\, \hat{d}(\mathbf{u}) + \xi - \tilde{d}(\mathbf{u}))^2,$$

where Ω denotes the set of valid pixels. The aligned depth is $\hat{D}(\mathbf{u}) = (\beta^* \hat{d}(\mathbf{u}) + \xi^*)^{-1}$, and the depth loss penalizes residuals using a robust Huber penalty $\mathcal{H}_\delta$ together with edge-aware smoothness:

$$\mathcal{L}_d = \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} \mathcal{H}_\delta(\epsilon(\mathbf{u})) + \frac{1}{2}(|\nabla_x \epsilon| + |\nabla_y \epsilon|) + \|\beta^* - 1\|_2^2 + \|\xi^*\|_2^2,$$

where $\epsilon(\mathbf{u}) = |\hat{D}(\mathbf{u}) - \tilde{D}(\mathbf{u})|$ and the penalties on β^* and ξ^* gently bias alignment toward identity. Similarly, the normal consistency loss is

$$\mathcal{L}_n = \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} (1 - \langle \hat{\mathbf{n}}(\mathbf{u}), \tilde{\mathbf{n}}(\mathbf{u}) \rangle).$$

Finally, to prevent view-dependent effects from leaking into albedo, we enforce Laplacian smoothness:

$$\mathcal{L}_\rho = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{N}(k)|} \sum_{l \in \mathcal{N}(k)} \|\boldsymbol{\rho}^{(k)} - \boldsymbol{\rho}^{(l)}\|_2^2, \tag{1}$$

The total objective is: $\mathcal{L} = \mathcal{L}_i + \lambda_d \mathcal{L}_d + \lambda_n \mathcal{L}_n + \lambda_\rho \mathcal{L}_\rho$.

4. Experiments

We evaluate GutSee on phantom colonoscopy data and compare it against existing splatting approaches. To ensure fairness, we test baselines under multiple supervision protocols and report both geometric and image-based reconstruction metrics.

Datasets. We benchmark GutSee on 5 phantom colonoscopy sequences sampled from C3VDv1 (Bobrow et al., 2023) and 5 from C3VDv2 (Golhar et al., 2025). To our knowledge, no 3D reconstruction approaches in the literature have evaluated their methods on C3VDv2, which introduces more challenging scenes characterized by complex camera trajectories and the presence of debris.

Baselines. As a general baseline, we consider 3DGS. For domain-specific baselines, we evaluate EndoGSLAM (Wang et al., 2024b), Gaussian Pancakes (Bonilla et al., 2024), and PR-ENDO (Kaleta et al., 2025). For all comparisons, we fix camera poses to their true values to isolate mapping quality from tracking performance. To ensure a comprehensive comparison, we evaluate all models under two configurations: (i) a ground truth-supervised setting, and (ii) a depth prior-supervised setting using ColonCrafter predicted depth maps. This two-tier evaluation disentangles the contributions of supervision signals and model architecture, and highlights GutSee's ability to operate robustly even with biased priors.

Metrics. We assess reconstruction quality using both geometry- and image-based metrics. For geometric accuracy, we compute the bidirectional Chamfer distance (CD) between reconstructed and ground truth point clouds to assess global surface alignment, and the depth root-mean-square error (D-RMSE) against reference depths to quantify metric depth accuracy. For image fidelity, we report peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS). Following standard practice, evaluations are performed on every eighth held-out frame. All metrics are computed per sequence and subsequently averaged across sequences.

Table 1: Quantitative comparison of GutSee with baseline methods on 10 sequences from C3VD (aggregated over v1 and v2). GutSee achieves the strongest geometric accuracy with biased depth supervision, while maintaining competitive rendering quality. Sup.: Supervision; GT: Ground Truth; CC: ColonCrafter.

	Method	CD ↓	D-RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GT Sup.	3DGS (Kerbl et al., 2023)	0.940	14.214	32.301	0.850	0.243
	EndoGSLAM (Wang et al., 2024b)	0.565	3.095	20.759	0.750	0.468
	Gaussian Pancakes (Bonilla et al., 2024)	0.478	7.939	30.602	0.829	0.296
	PR-ENDO (Kaleta et al., 2025)	0.620	0.946	31.706	0.843	0.274
	GutSee (ours)	0.498	1.161	30.025	0.826	0.266
CC Sup.	3DGS (Kerbl et al., 2023)	2.154	15.112	33.425	0.882	0.219
	EndoGSLAM (Wang et al., 2024b)	3.049	8.576	20.279	0.717	0.481
	Gaussian Pancakes (Bonilla et al., 2024)	2.249	8.449	31.269	0.860	0.276
	PR-ENDO (Kaleta et al., 2025)	2.583	5.452	31.992	0.864	0.268
	GutSee (ours)	1.636	4.452	30.392	0.855	0.250

🟥 Worst   🟨 Middle   🟩 Best

Implementation Details. Across experiments, all frames are undistorted and resized to 384×384 pixels, with camera intrinsics scaled accordingly. We optimize each scene for 15,000 iterations. Densification and pruning follows the logic of the original Triangle Splatting implementation; we prune and densify triangles every 500 iterations, starting from iteration 500 and stopping at iteration 13,000. We set the initial learning rate of the triangle positions to 2×10^{-3}, and choose $\lambda_{\text{SSIM}} = 0.2$, $\lambda_d = 0.1$, $\lambda_n = 0.1$, and $\lambda_\rho = 0.1$ for the loss weights. All models are trained using PyTorch (Paszke et al., 2019) on a single NVIDIA A10 Tensor Core GPU.

Results. The results in Table 1 highlight two central findings. First, GutSee delivers consistently strong geometric reconstruction accuracy. With ground truth supervision, it outperforms EndoGSLAM and 3DGS but lags slightly behind Gaussian Pancakes and PR-ENDO; we hypothesize that this is because both methods incorporate a loss term that explicitly encourages the scene representation not to stray from the initial scene (which in this case is the ground truth). Under biased ColonCrafter supervision, GutSee yields substantially lower geometric error than all baselines (1.636 mm CD, 4.452 mm D-RMSE); the next best performing model, PR-ENDO, has an average D-RMSE 1.000 mm greater. Notably, GutSee maintains acceptable rendering quality with an average LPIPS value of 0.250, second only to 3DGS which heavily overfits the images; as shown in Figure 3, the depth renders produced by 3DGS are wholly unfaithful to the true geometry of the colon. Together, these results indicate that GutSee offers a robust geometry-centric reconstruction approach that does not compromise rendering quality.

Figure 3 further provides a qualitative comparison of novel-view renderings across methods on the `cecum_t1_a` sequence from C3VD. While Gaussian Pancakes and 3DGS recover diffuse details well, their renders exhibit muted or spatially imprecise specular highlights. PR-ENDO models illumination through an MLP, enabling more faithful specular effects, yet it can struggle with fine details and produces comparatively noisy depth maps. In contrast, GutSee produces smooth surface reconstructions while simultaneously being able to repro-
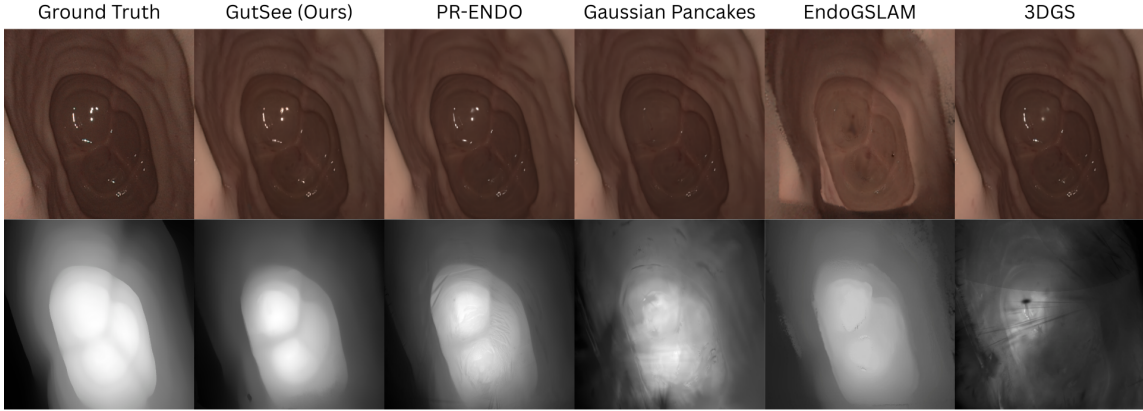
Figure 3: Qualitative comparison of novel-view renderings and depth estimates on the cecum_t1_a sequence. GutSee produces illumination-consistent renderings and depth maps with reduced artifacts and sharper geometric delineation relative to competing approaches.

duce complex endoscopic lighting phenomena, even when trained under biased ColonCrafter priors.

Another advantage of our approach lies in its explicit separation of lighting and surface geometry. Because our formulation employs a physically grounded BRDF parameterization with learnable materials and an explicit light source, the method can successfully disentangle illumination effects from underlying shape. As shown in Figure 4, specular highlights emerge directly from the interaction between the estimated light source and the reconstructed surface, rather than being absorbed into material parameters or inadvertently encoded in geometry.

5. Discussion and Conclusion

In this work, we introduced GutSee, a depth-guided triangle splatting framework for monocular colonoscopic reconstruction that combines affine-invariant geometric supervision with an explicit model of endoscopic illumination. On phantom colonoscopy sequences, GutSee achieves lower Chamfer distance and depth RMSE than splatting-based baselines under biased supervision, while maintaining competitive photometric performance (Table 1). These results indicate that explicitly addressing depth ambiguity and light-geometry entanglement is an effective strategy for geometry-centric reconstruction in endoscopy.

However, defining what constitutes a "realistic" colonoscopic reconstruction remains non-trivial. Global geometric metrics such as Chamfer distance or depth RMSE capture overall alignment but can hide localized errors in folds, narrow segments, or regions of strong specularities. Conversely, image-based metrics may favor visually plausible but geometrically inconsistent solutions. Although GutSee improves geometric accuracy relative to baselines, subtle deviations that could matter clinically (such as local underestimation of fold height or lesion protrusion) may not be fully captured by the current evaluation
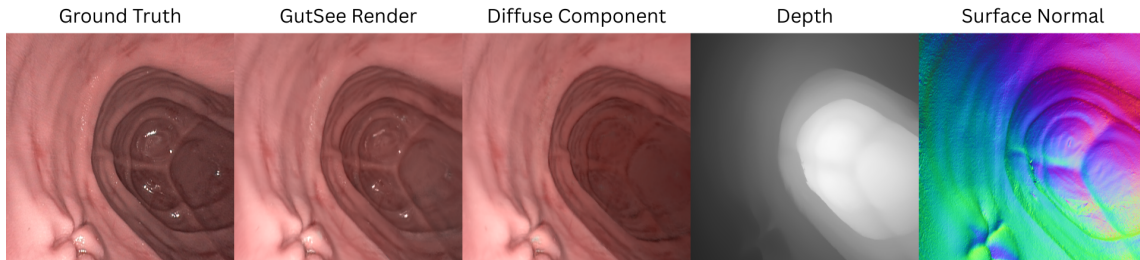
Figure 4: An example of light-geometry decoupling. Our physically based scene representation and explicit coaxial light model separates lighting effects from the underlying scene geometry.

protocol. As in other generative and reconstruction settings, the ultimate quality of the reconstructions should be judged by how well they support downstream tasks, including navigation, lesion localization, and quantitative measurement.

Several limitations suggest directions for future work. Our quantitative evaluation is restricted to phantom sequences and does not cover the full spectrum of in vivo variability, including strong non-rigid motion, fluids, and patient-specific variation. Moreover, GutSee remains coupled to the quality of the depth prior: although the affine-invariant loss prevents incorrect local affine biases from being hard-coded into the reconstruction, it does not "repair" fundamentally inaccurate priors, and severe local errors or missing structures in the ColonCrafter predictions can still lead the optimizer to suboptimal geometry. In addition, the current formulation assumes access to known camera poses; performing full SLAM with jointly optimized poses and geometry over extended, realistic trajectories remains challenging. Extending GutSee to real clinical videos, enriching the illumination and material models, and learning transferrable features across scenes are promising directions for future work.

In summary, GutSee provides a practical framework for physically consistent, geometry-aware reconstruction in colonoscopy using only monocular RGB video and pose information. By combining affine-invariant depth guidance with explicit triangle-based geometry and a coaxial lighting model, it moves reconstruction methods closer to being not only visually convincing but also geometrically reliable enough to support computer-assisted colonoscopy.

References

Mohamed Abdelrahim, Hiroyasu Saiga, Naoto Maeda, Ejaz Hossain, Hitoshi Ikeda, and Pradeep Bhandari. Automated sizing of colorectal polyps using computer vision. *Gut*, 71 (1):7–9, 2022.

Bradley W Anderson, Thomas C Smyrk, Kari S Anderson, Douglas W Mahoney, Mary E Devens, Seth R Sweetser, John B Kisiel, and David A Ahlquist. Endoscopic overestimation of colorectal polyp size. *Gastrointestinal endoscopy*, 83(1):201–208, 2016.

Víctor M Batlle, José MM Montiel, Pascal Fua, and Juan D Tardós. Lightneus: Neural surface reconstruction in endoscopy using illumination decline. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 502–512. Springer, 2023.

Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S Akshintala, Juan R Garcia, and Nicholas J Durr. Colonoscopy 3d video dataset with paired depth from 2d-3d registration. *Medical image analysis*, 90:102956, 2023.

Sierra Bonilla, Shuai Zhang, Dimitrios Psychogyios, Danail Stoyanov, Francisco Vasconcelos, and Sophia Bano. Gaussian pancakes: geometrically-regularized 3d gaussian splatting for realistic endoscopic reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 274–283. Springer, 2024.

Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm siggraph*, volume 2012, pages 1–7. vol. 2012, 2012.

Mayank V Golhar, Lucas Sebastian Galeano Fretes, Loren Ayers, Venkata S Akshintala, Taylor L Bobrow, and Nicholas J Durr. C3vdv2–colonoscopy 3d video dataset with enhanced realism. *arXiv preprint arXiv:2506.24074*, 2025.

John J Han and Jie Ying Wu. Endopbr: Material and lighting estimation for photorealistic surgical simulations via physically-based rendering. *arXiv preprint arXiv:2502.20669*, 2025.

Romain Hardy, Tyler Berzin, and Pranav Rajpurkar. Coloncrafter: A depth estimation model for colonoscopy videos using diffusion priors, 2025. URL https://arxiv.org/abs/2509.13525.

Jan Held, Renaud Vandeghen, Adrien Deliege, Abdullah Hamdi, Silvio Giancola, Anthony Cioppa, Andrea Vedaldi, Bernard Ghanem, Andrea Tagliasacchi, and Marc Van Droogenbroeck. Triangle splatting for real-time radiance field rendering. *arXiv preprint arXiv:2505.19175*, 2025a.

Jan Held, Renaud Vandeghen, Abdullah Hamdi, Adrien Deliege, Anthony Cioppa, Silvio Giancola, Andrea Vedaldi, Bernard Ghanem, and Marc Van Droogenbroeck. 3d convex splatting: Radiance field rendering with 3d smooth convexes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21360–21369, 2025b.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024.

Joanna Kaleta, Weronika Smolak-Dyżewska, Dawid Malarz, Diego Dall'Alba, Przemysław Korzeniowski, and Przemysław Spurek. Pr-endo: Physically based relightable gaussian splatting for endoscopy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 391–401. Springer, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Ruibin Ma, Rui Wang, Stephen Pizer, Julian Rosenman, Sarah K McGill, and Jan-Michael Frahm. Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 573–582. Springer, 2019.

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

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Sam A O'Connor, David G Hewett, Marcus O Watson, Bradley J Kendall, Luke F Hourigan, and Gerald Holtmann. Accuracy of polyp localization at colonoscopy. *Endoscopy international open*, 4(06):E642–E646, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Dimitrios Psychogyios, Francisco Vasconcelos, and Danail Stoyanov. Realistic endoscopic illumination modeling for nerf-based data generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 535–544. Springer, 2023.

Sara Pyykölä, Niclas Joswig, and Laura Ruotsalainen. Non-lambertian surfaces and their challenges for visual slam. *IEEE Open Journal of the Computer Society*, 5:430–445, 2024.

Douglas K Rex, C Richard Boland, Jason A Dominitz, Francis M Giardiello, David A Johnson, Tonya Kaltenbach, Theodore R Levin, David Lieberman, and Douglas J Robertson. Colorectal cancer screening: recommendations for physicians and patients from the us multi-society task force on colorectal cancer. *Gastroenterology*, 153(1):307–323, 2017.

Adam Schmidt, Omid Mohareri, Simon DiMaio, Michael C Yip, and Septimiu E Salcudean. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, 94:103131, 2024.

Robert E Schoen, Lawrence D Gerber, and Craig Margulies. The pathologic measurement of polyp size is preferable to the endoscopic estimate. *Gastrointestinal endoscopy*, 46(6):492–496, 1997.

Benjamin M Shandro, Khemraj Emrith, Gregory Slabaugh, Andrew Poullis, and Melvyn L Smith. Optical imaging technology in colonoscopy: Is there a role for photometric stereo? *World Journal of Gastrointestinal Endoscopy*, 12(5):138, 2020.

Domenico Soriero, Paola Batistotti, Rafaela Malinaric, Davide Pertile, Andrea Massobrio, Lorenzo Epis, Beatrice Sperotto, Veronica Penza, Leonardo S Mattos, Marina Sartini,

et al. Efficacy of high-resolution preoperative 3d reconstructions for lesion localization in oncological colorectal surgery—first pilot study. In *Healthcare*, volume 10, page 900. MDPI, 2022.

Atalie C Thompson, Richard H Jones, Peter D Poullos, Subhas Banerjee, and Lewis K Shin. Taller haustral folds in the proximal colon: A potential factor contributing to interval colorectal cancer. *Journal of Colon and Rectal Cancer*, 1(1):45–54, 2016.

Jeroen C Van Rijn, Johannes B Reitsma, Jaap Stoker, Patrick M Bossuyt, Sander J Van Deventer, and Evelien Dekker. Polyp miss rate determined by tandem colonoscopy: a systematic review. *Official journal of the American College of Gastroenterology— ACG*, 101 (2):343–350, 2006.

Jiepeng Wang, Yuan Liu, Peng Wang, Cheng Lin, Junhui Hou, Xin Li, Taku Komura, and Wenping Wang. Gaussurf: Geometry-guided 3d gaussian splatting for surface reconstruction. *arXiv preprint arXiv:2411.19454*, 2024a.

Kailing Wang, Chen Yang, Yuehao Wang, Sikuang Li, Yan Wang, Qi Dou, Xiaokang Yang, and Wei Shen. Endogslam: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 219–229. Springer, 2024b.

Xiongzhi Wang, Boyu Yang, Min Wei, Liangfa Liu, Jingang Zhang, and Yunfeng Nie. Deep learning for endoscopic depth estimation: A review. *Displays*, page 103086, 2025.

Zhongyi Wang, Mengjiao Shen, and Qijun Chen. Eliminating scale ambiguity of unsupervised monocular visual odometry. *Neural Processing Letters*, 55(7):9743–9764, 2023.

Xinkai Yan, Jieting Xu, Yuchi Huo, and Hujun Bao. Neural rendering and its hardware acceleration: A review. *arXiv preprint arXiv:2402.00028*, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

Heming Yao, Ryan W Stidham, Zijun Gao, Jonathan Gryak, and Kayvan Najarian. Motion-based camera localization system in colonoscopy videos. *Medical image analysis*, 73: 102180, 2021.

Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf: Neural incident light field for physically-based material estimation. In *European conference on computer vision*, pages 700–716. Springer, 2022.

Linglong Zhou, Guoxin Wu, Yunbo Zuo, Xuanyu Chen, and Hongle Hu. A comprehensive review of vision-based 3d reconstruction methods. *Sensors*, 24(7):2314, 2024.

Appendix A. Visualization of Learned Albedo Parameters and Geometry

Figure 5 visualizes the diffuse albedo inferred by GutSee. The model recovers material properties that are disentangled from scene illumination, yielding intrinsic parameters that support physically consistent reconstruction and rendering.
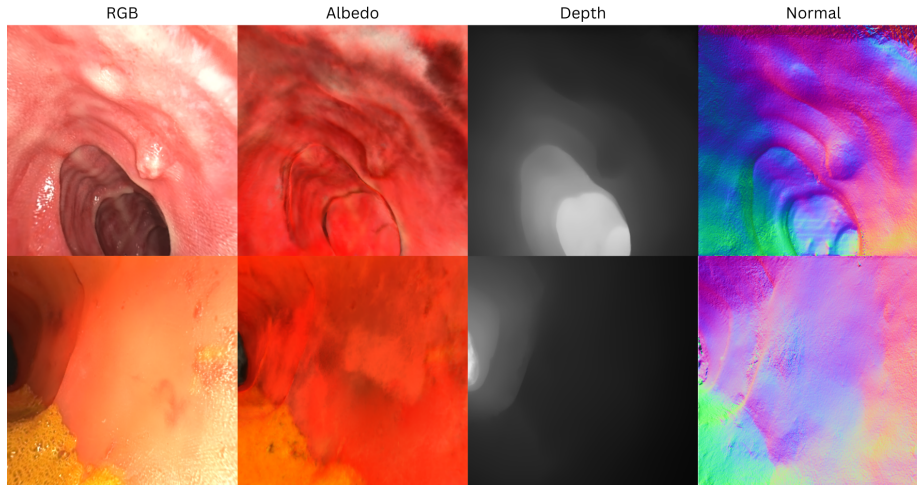

Figure 5: Visualization of intrinsic material parameters estimated by GutSee. From left to right: held-out rendered RGB, recovered diffuse albedo, predicted depth, and estimated surface normals. The method learns illumination-invariant albedo and geometrically consistent structure from posed colonoscopy frames and biased monocular depth priors.

Appendix B. Novel View Synthesis on Additional C3VD Sequences

Figure 6 presents supplementary qualitative comparisons of novel view synthesis on a diverse subset of C3VD scenes. Across all examples, GutSee reconstructs surfaces with high geometric fidelity and delivers view-consistent appearance, while maintaining photometric quality comparable to existing baselines.
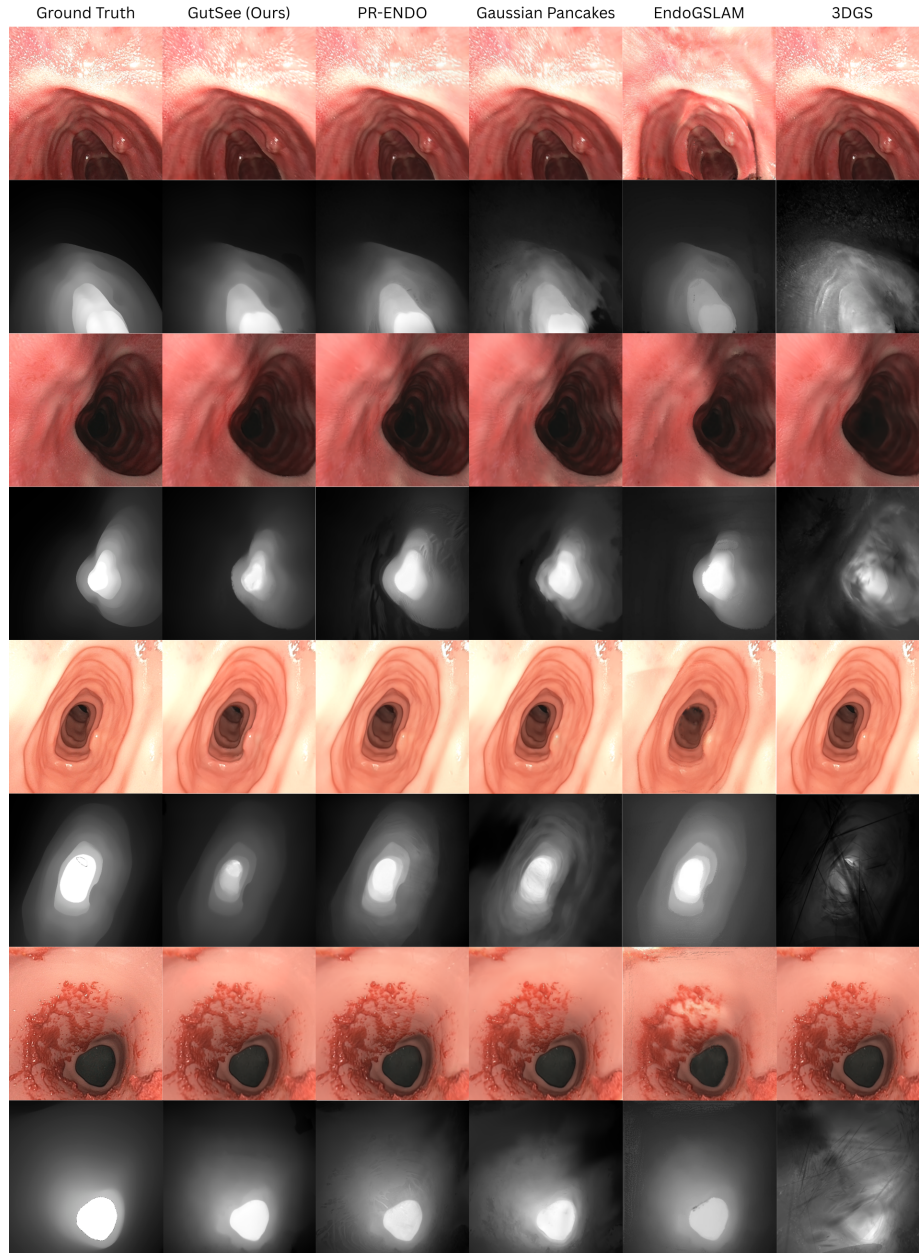
Figure 6: Additional novel-view synthesis results on representative C3VD sequences. Gut-See reconstructs geometrically coherent and well-localized surfaces across varied anatomical structures, while achieving photometric accuracy at least on par with neural rendering baselines.

Appendix C. Ablation Study: Loss Terms

We perform a light ablation study on the `cecum_t1_a` sequence to evaluate the effect of $\mathcal{L}_d$, $\mathcal{L}_n$, and our affine-invariant formulation on reconstruction performance. Table 2 lists geometric and image reconstruction metrics for this scene under (i) our base configuration, (ii) without explicit geometric supervision ($\lambda_d = \lambda_n = 0$), and (iii) with geometric supervision but no affine alignment step (i.e., we use the ColonCrafter priors to supervise optimization directly). Overall, our base configuration produces the most desirable results. Although CD decreases and photometric rendering performance improves without geometric supervision, D-RMSE increases by 1.975 mm and the model no longer renders smooth depths. Without affine alignment, we find that both CD and D-RMSE worsen, suggesting that the model inherits bias from the ColonCrafter priors.

Table 2: Ablation study on the use of geometric loss terms and our affine alignment step. Our base configuration produces the best overall results, motivating the use of affine-invariant geometric guidance.

Training configuration	CD ↓	D-RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
(i) Base configuration	2.057	5.324	32.301	0.929	0.277
(ii) No geometric supervision ($\lambda_d = \lambda_n = 0$)	1.804	7.299	35.756	0.936	0.266
(iii) No affine alignment	2.358	5.753	34.078	0.927	0.278