Leveraging Near-Field Lighting for Monocular Depth Estimation from Endoscopy Videos

Akshay Paruchuri[†], Samuel Ehrenstein, Shuxian Wang, Inbar Fried, Stephen M. Pizer, Marc Niethammer, and Roni Sengupta[†]

Department of Computer Science
University of North Carolina at Chapel Hill

Abstract. Monocular depth estimation in endoscopy videos can enable assistive and robotic surgery to obtain better coverage of the organ and detection of various health issues. Despite promising progress on mainstream, natural image depth estimation, techniques perform poorly on endoscopy images due to a lack of strong geometric features and challenging illumination effects. In this paper, we utilize the photometric cues, i.e., the light emitted from an endoscope and reflected by the surface, to improve monocular depth estimation. We first create two novel loss functions with supervised and self-supervised variants that utilize a per-pixel shading representation. We then propose a novel depth refinement network (PPSNet) that leverages the same per-pixel shading representation. Finally, we introduce teacher-student transfer learning to produce better depth maps from both synthetic data with supervision and clinical data with self-supervision. We achieve state-of-the-art results on the C3VD dataset while estimating high-quality depth maps from clinical data. Our code, pre-trained models, and supplementary materials can be found on our project page: `https://ppsnet.github.io/`.

Keywords: Monocular depth estimation · Photometric refinement · Sim2Real transfer learning · Endoscopic imaging

1 Introduction

Endoscopy [7] is a minimally invasive imaging technique where a slender and flexible tube equipped with a light and a camera is used to inspect internal organs and tissues of the body via an opening, e.g., the mouth or anus. Endoscopy is routinely used for cancer screening, e.g., in colons [28] and lungs [1]; for identifying potential issues in the digestive tract, e.g., ulcers, polyps, and hemorrhoids [22]; and for respiratory problems, e.g., infections, abnormal narrowing of the airway, and collapsed lungs [12]. In 2022 an estimated 225.3 million endoscopic procedures were performed globally with an annual growth rate of 1.3% [17]. Endoscopy is challenging as it requires highly skilled personnel navigating through narrow and winding pathways in the body. As a result, patients

---

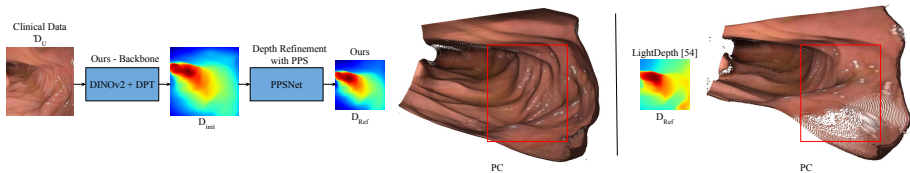[†]Corresponding authors: {akshay, ronisen}@cs.unc.edu

Fig. 1: Our approach models near-field lighting, emitted by the endoscope and reflected by the surface, as Per-Pixel Shading (PPS). We use PPS feature to perform depth refinement (PPSNet) on clinical data using teacher-student transfer learning and a PPS-informed self-supervision. Our method outperforms the state-of-the-art monocular depth estimation technique LightDepth [54], which trains using self-supervision that explicitly reconstructs input images using illumination decline.

may experience discomfort and some regions may remain unsurveyed. Thus researchers have focused on developing intelligent techniques for better guidance and visualization of unsurveyed regions during the procedure, to obtain better coverage of the organ, and to facilitate autonomous navigation. However, these intelligent techniques are still clinically infeasible as they all share the fundamental problem of accurately estimating the 3D organ geometry. 3D understanding of organ shapes will also enable automatic measurement of geometric properties, e.g., measuring the airway cross-sectional area in children to detect any abnormalities without exposing them to radiation [18].

Current attempts [5,23,41,42,52,61,65,70,72] at 3D understanding of internal organs from endoscopy videos heavily rely on learning priors from synthetic training data and leveraging geometric cues from camera motion. While this leads to a reasonable 3D understanding on synthetic data [8, 52] and imagery captured with a cellphone camera in typical outdoor and indoor environments [24, 67], it often fails on real clinical data due to the lack of strong geometric features, strong textural cues, strong reflections from mucus layers, and complex illumination effects. This issue is even more pronounced for oblique and en-face views where the camera is looking away from the tubular structures of the colon or the airway, which is often crucial for detecting cancer and other health screenings.

Our key idea is to model the reflection of co-located light with the camera of the endoscope from the surface of the internal organ to help predict the relative distance between the endoscope and the surface. We rely on the fact that any point on the surface closer to the camera and facing the endoscope will reflect more light than a point that is further away or facing away from the endoscope. Thus we introduce a monocular depth estimation algorithm that can leverage near-field illumination and learn from synthetic data with ground-truth supervision and real clinical data with self-supervision. While our method is general for any endoscopy videos, in this work we focus on colonoscopy. This is because colonoscopy screening [43] is the gold standard for colorectal cancer detection, the third most common cancer in the world, and because datasets of both synthetic (phantom [8] and computer graphics generated [52]) and real colonoscopies [2, 3, 40] are publicly available.

We propose multiple key contributions that harness the power of near-field illumination in predicting depth information. ● We first utilize the per-pixel lighting representation, introduced in [32], to compute the Per-Pixel Shading (PPS) map which is strongly correlated with the image intensity field. In contrast to [32] which only uses per-pixel lighting representation as an input to a surface normal estimation network, we calculate a PPS map and propose a supervised and a self-supervised loss function for training on synthetic and real data respectively. ● We then propose a depth refinement architecture, PPSNet, that modulates depth features computed by a monocular depth estimation foundation model, combining DINOv2 [44] features and an initial depth from a DPT-based [50] decoder, fine-tuned on the C3VD [8] dataset, with features computed from the PPS map. To our knowledge, existing depth refinement networks [4, 29, 33] often rely on geometric consistency or uncertainty estimation, whereas we exploit illumination features which are more useful in endoscopy videos. ● Finally, we propose a teacher-student transfer learning approach, where the teacher network, effectively a monocular depth estimation foundation model followed by PPSNet refinement, is trained on both synthetic data with supervision and real clinical data with weak supervision. The Teacher additionally guides the learning of the Student network on unlabeled real clinical data with our proposed self-supervised loss function that utilizes PPS representation. In contrast to teacher-student learning-based monocular depth estimation algorithm [67] that trains on ∼60 million unlabeled images with semantic self-supervision, our method is trained with only ∼18,000 clinical endoscopy images using lighting based self-supervision.

We show that our proposed approach significantly outperforms state-of-the-art monocular depth estimation techniques on the synthetic C3VD dataset [8] both quantitatively and qualitatively. Similar observations also hold true for real clinical datasets [2,40], which we show qualitatively in the absence of any ground-truth depth maps. We perform a detailed ablation study to show the effectiveness of our per-pixel shading-based loss functions, depth refinement modules, and our teacher-student-based transfer learning approach.

In summary, the key contributions of our approach are: ● We introduce a supervised and a self-supervised loss function that utilizes near-field illumination emitted by the endoscope and reflected by the surface. ● We propose a depth-refinement architecture, PPSNet, that utilizes near-field illumination. ● We develop a teacher-student transfer learning approach where a teacher model guides the learning of a student model using our proposed self-supervised loss function. ● We perform an extensive evaluation on synthetic C3VD [8] and real clinical data [2, 40], producing state-of-the-art results. We will release all codes, data, and benchmarks for future research.

2   Related Works

Monocular Depth Estimation. Monocular depth estimation has received a great deal of attention due to its importance in 3D understanding from a single image. Researchers have considered various different approaches ranging from

hand-crafted features [20, 34, 60], convolutional neural networks [9, 11, 13, 31, 49, 51, 66], to transformers [50]. The success of transformer-based techniques also prompted larger-scale data-driven approaches that mixed datasets [10] and utilized synthetic data to better generalize to unseen data [24]. Recently, Depth Anything [67] showed great monocular depth estimation quality by utilizing both large-scale labeled and unlabeled data.

However, most of these advances focus on natural images captured outdoors or indoors with cellphone cameras. These methods typically do not generalize well to endoscopy videos due to a lack of strong geometric features, challenging illumination conditions, and strong reflections from mucus layers. Until recently, endoscopy data with ground-truth depth maps was difficult to obtain. Thus self-supervised or unsupervised techniques utilizing limited data and techniques such as weak depth supervision [21,35,39] and uncertainty estimation [55] showed promise. Multi-view approaches [39] combined with SLAM pipelines that integrate tracking and depth estimation [41, 45, 53] further showed promise despite being encumbered by weak texture cues, limited degrees of camera motion, and the presence of organ tissue deformations. Recent efforts in creating large-scale synthetic datasets of colonoscopy videos by using computer graphics engines (SimCol3D [52]) or phantom models (C3VD [8]), as well as unlabeled datasets composed of clinical sequences, e.g., Colon10K [40] and EndoMapper [3], improved monocular depth estimation by utilizing both synthetic and real data [64]. However, these approaches often fail to generalize to clinical data, performing extremely badly on non-axial views (comprising 40-50% of the data), where the camera is looking away from the tubular colon structure, due to lack of strong geometric priors. In this paper, we propose a monocular depth estimation algorithm that leverages illumination information along with geometric features.

Depth Using Illumination Information. Researchers have developed techniques where one can reconstruct a surface from multiple images captured with varying illumination conditions, also called Photometric Stereo [6,33,37,68,71]. One special focus is near-field Photometric Stereo, where the light source is close to the object resulting in different points on the surface receiving different intensities of light. Numerous approaches to near-field photometric stereo utilize a per-pixel lighting representation to effectively model scene lighting and for subsequent estimation or refinement of surface normals in an optimization framework [36, 46, 59] or as an input to a neural network [32]. We also use per-pixel lighting information, but go two steps forward, to create a supervised and self-supervised loss function that enables us to train on unlabeled real clinical data. Additionally, we create a depth refinement network that uses a per-pixel shading map to modulate an initial depth prediction. In contrast to previous near-field photometric stereo approaches we only require a single image captured with a co-located light and camera.

Recently, LightDepth [54] also proposed self-supervised loss functions that used illumination attenuation with differentiable rendering for image reconstruction. Our proposed self-supervised loss operates on per-pixel shading representation and avoids explicit rendering of an image with imperfect assumptiosn of
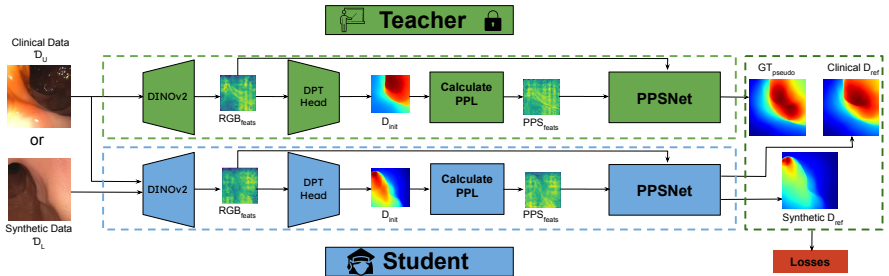
Fig. 2: An overview of our proposed approach to train a student network capable of producing high quality depth maps on both synthetic colonoscopy data and real clinical colonoscopy data. For simplicity, $PPSNet$ here corresponds to lines 4-7 of alg. 1.

diffuse albedo and direct lighting in LightDepth [54]. We also go beyond self-supervised loss, by developing supervised per-pixel shading loss, teacher-student transfer learning and illumination-aware depth refinement module, producing significantly better result than LightDepth on both synthetic and clinical data.

Depth Refinement Network. Depth refinement is a commonly employed technique that leverages an initial predicted depth and additional information, such as surface normals, multi-scale information, or depth context, in order to produce higher quality depth maps [25, 30, 58]. Recent approaches utilize geometric consistency, uncertainty estimation, or guidance masks to refine depths [4, 26, 29]. We propose a depth refinement module that modulates an initial depth prediction with extracted RGB features and features extracted from Per-Pixel Shading, which is calculated from the initial depth prediction, to generate the refined depth - a first to our knowledge.

Teacher-Student Transfer Learning. Teacher-student transfer learning is a popular form of knowledge distillation [14, 16, 19, 56, 69]. Depth Anything [67], a monocular depth estimation technique, utilizes teacher-student learning while leveraging large-scale data (1.5 million labeled images, 62 million unlabeled images) and training with a self-supervised loss function that includes semantic losses. For endoscopy, such large-scale data is unavailable and semantic losses are not meaningful. Our proposed teacher-student learning uses a lighting-aware self-supervised loss function using only $\sim 18{,}000$ images.

3   Our Approach

Our approach utilizes per-pixel-shading (PPS) fields as a part of our learning objectives and refinement module to produce a model capable of achieving the state-of-the-art monocular depth estimation on the synthetic C3VD [8] colonoscopy dataset. In the following sections, we describe the analytical computation of per-pixel lighting, the PPS field in Sec. 3.1, and its subsequent use in both our learning objectives (Sec. 3.2) and as an input to our novel depth refinement module (Sec. 3.3). We then propose a teacher-student transfer learning that learns from unlabeled, real clinical data using our proposed self-supervised loss function, described in Sec. 3.4.

3.1   Per-Pixel Lighting Representation

Existing depth estimation techniques [24, 50, 51, 67] rely on learning geometric and semantic priors from their training data. However, these methods ignore crucial photometric information that can be obtained from a near-field co-located light and camera. The surfaces of an internal organ closer to the camera and facing the camera receive more incident illumination than surfaces further from the camera or facing in opposite directions. Our key idea is to leverage this fact by modeling incident illumination at every point on the surface using a Per-Pixel Lighting (PPL) representation, following [32]. However, our method improves upon Lichy *at al.* [32], which only uses per-pixel lighting as an input to a neural network for surface normal estimation, by explicitly designing loss functions to train on synthetic and unlabeled real data and by designing a depth refinement architecture using feature modulation. In this paper, we focus on colons due to publicly available datasets, but our method is general and can be extended to other organs like the upper airway.

Camera Model. We utilize the conventional pinhole camera framework, positioned at the origin in the global coordinate system and looking down the z-axis. This camera is defined by the intrinsic matrix K. A point in the world space, denoted as $X = (x, y, z)$, is mapped to a pixel coordinate (u, v) using: $(u, v, 1)^T \sim K(x, y, z)^T$.

Geometry Model. We only take into account visible surfaces of the colon, and therefore assume that the depth map and corresponding normal map is a complete description of in-view surfaces. $X(u, v) \in \mathbb{R}^3$ describes the 3D location of a pixel (u, v) on the surface of the organ. The depth map can then be defined by $D(u, v) = X(u, v)_3$, corresponding to the z-component of $X(u, v)$. $X(u, v)$ itself can be recovered from the following: $X(u, v) = D(u, v)K^{-1}(u, v, 1)^T$.

Per-Pixel Lighting. Following the notations introduced in [32], we represent any point light source, i.e. the co-located light of the endoscope, with its location $p \in \mathbb{R}^3$, light direction $d \in \mathbb{S}^2$, and angular attenuation coefficient $\mu \in \mathbb{R}$. Then for any point X on the surface of the colon, we can represent Per-Pixel Lighting (PPL) using a lighting direction $L(X)$ and an attenuation factor $A(X)$ as follows:

$$L(X) = \frac{X - p}{\|X - p\|}, \quad A(X) = \frac{(L \cdot d)^\mu}{\|X - p\|^2}. \tag{1}$$

Thus we can analytically compute PPL information $L(u, v)$ and $A(u, v)$ for every pixel (u, v) of an image captured with a co-located light source and camera, i.e., an endoscope, given the depth map $D(u, v)$ and the camera intrinsics K. This presents a chicken-and-egg problem where the PPL information is strongly correlated with depth, i.e. one can be computed from the other, despite both being unknown initially. Our proposed solution to this problem is to leverage ground truth and pseudo-ground truth where possible, as well as a self-supervised loss objective that takes advantage of the aforementioned strong correlation. For our calculation of per-pixel lighting (PPL) information, we assume the camera and light source are co-located with the camera direction being in the positive

z direction. Additionally, we assume that our light is modeled by an isotropic point source. Therefore, the angular attenuation coefficient μ is 0 (rendering d in the calculation of $A(X)$ in eqn. 1 irrelevant).

3.2 Proposed Loss Functions Using Per-Pixel Lighting

Following the PPL representation, introduced in [32] and summarized in Sec. 3.1, we propose two new loss functions, a supervised loss function for training on synthetic data, and a self-supervised loss function for training on clinical data.

First, we propose a new representation, the Per-Pixel Shading field PPS, that combines the analytically computed PPL information of lighting direction, $L(X)$, and attenuation factor, $A(X)$, with the surface normal $N(X)$ to capture the effects of incident lighting on the surface geometry:

$$PPS(X) = A(X) \times (L(X) \cdot N(X)). \tag{2}$$

The surface normal can be computed by differentiating the depth map $D(u, v)$. A visual illustration of our proposed PPS field can be found in Fig. 3. PPS captures the diffuse shading effect, which indicates how light emitted from the endoscope will be reflected by the surface, ignoring strong specularities and inter-reflections.

Supervised PPS Loss. We introduce a loss function that computes the mean squared error between the ground-truth Per-Pixel Shading (PPS), computed using the ground truth depths and surface normals, and the predicted PPS, analytically computed from a final, refined depth estimate. For an image of resolution $H \times W$, we can compute this loss using eqn. 3 where specularities are masked with mask function M, s.t. $M(u, v) = 1$, $if\ I_g < 0.98$, where I_g is the intensity map of the input image. We believe that the supervised PPS loss is complementary to supervised depth loss due to the inherent smoothness and low-frequency bias in the depth maps compared to more high-frequency information in the PPS map. Additionally, the lighting information in the PPS map can also help to better resolve distance ambiguity using the inverse squared falloff. Our experimental analysis in Tab. 2 also shows that on the synthetic C3VD dataset using $\mathcal{L}_{PPS-sup}$ helps to improve depth prediction.

$$\mathcal{L}_{PPS-sup} = \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} M(u, v) \left(PPS_{pred}(u, v) - PPS_{gt}(u, v)\right)^2. \tag{3}$$

Self-supervised PPS Loss. We introduce an additional loss function that utilizes the Per-Pixel Shading (PPS) representation but does not require the knowledge of ground-truth depths and surface normals and thus can be used for self-supervised learning. Our idea relies on the fact that the PPS is strongly correlated with the image intensity field except in regions of strong specularity and it ignores inter-reflections by only modeling direct, in-view illumination from the surface to the camera. We show this correlation in Fig. 3, and we observe that on the synthetic C3VD dataset that utilizes a phantom colon mold, the ground-truth PPS has a mean correlation of 0.90 with a variance 0.03 with the image intensity field. This high correlation indicates that the model predictions

for PPS are closely aligned with changes in the image intensity field, suggesting consistent and accurate modeling. The low variance points towards the stability of this relationship across different samples, meaning that our approach's use of the PPS is uniform and dependable across the entire dataset.
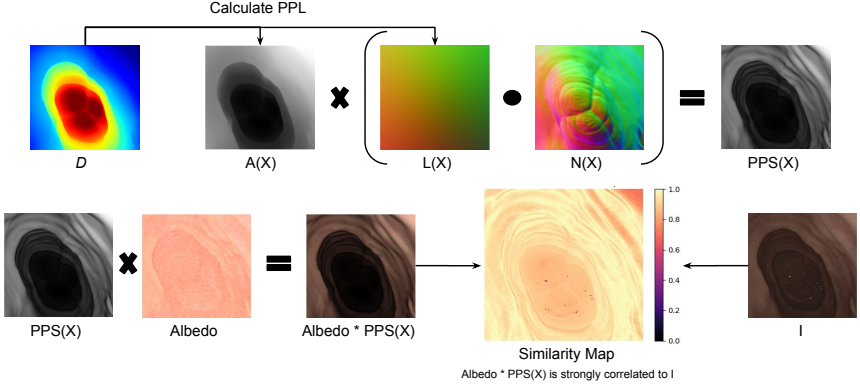

Fig. 3: An example of how we compute our PPS representation using depths and surface normals. The computed albedo-modulated PPS representation is strongly correlated to the corresponding input image.

Based on this observation, we introduce a self-supervised loss function that computes the correlation between the PPS, derived from the predicted depth map, and the grayscale image (intensity field) I_g, where specularities are masked with mask function M using a threshold of 0.98:

$$\mathcal{L}_{PPS-corr} = 1 - \text{corr}(M * I_g, M * PPS_{pred}). \tag{4}$$

The self-supervised loss function will enable us to train on real clinical data where ground-truth depth information is unavailable. Further details regarding the calculation of $\mathcal{L}_{PPS-corr}$ can be found in our supplementary materials.

3.3   Depth Refinement Module

We propose to use illumination information, represented by Per-Pixel Shading to enhance the depth refinement process. Our depth refinement module, called PPSNet, takes in RGB features (RGB_{feats}), PPS features (PPS_{feats}), and an initial depth prediction (D_{init}) in order to predict a value Δ_D to refine D_{init} with. All features are extracted using a DINOv2 [44] encoder. The initial depth prediction is obtained from a DPT-based decoder as described in [67]. Since DINOv2 expects an RGB image, we compute PPS features by simply multiplying the calculated Per-Pixel Shading (PPS) map with a proxy albedo, effectively an H, S value pair extracted from the input RGB image with V set to 100%. Due to the colon surfaces having a small variance in albedo, it is reasonable to create and use such a proxy for albedo. Integration of these features is achieved through a multi-headed cross-attention [63], which can be simply viewed as eqn. 5:

$$x_{combo} = \text{CrossAttention}(Q, K) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{5}$$

where $Q = RGB_{feats}$, $K = V = PPS_{feats}$, d_k denotes the dimensionality of the keys, and x_{combo} represents the synthesized feature set. These combined features influence D_{init} via feature-wise linear modulation (FiLM [48]), altering it to D_{mod} in eqn. 6:

$$D_{mod} = \gamma(x_{combo}) \cdot D_{init} + \beta(x_{combo}), \quad (6)$$

where γ and β are the scale and shift parameters generated by the FiLM module, dependent on x_{combo}.

Following modulation, D_{mod} is fed into a four-layer UNet, producing Δ_D, which is then finally added to D_{init} to yield D_{ref}. This sequence of steps describes our approach to using PPS for refinement and is further detailed within algorithm 1, which takes in the input image I, light position p, light direction d, angular attenuation coefficient μ, and camera intrinsics K. Additional insights into our approach are available in the supplementary materials.

Algorithm 1 A forward pass of our approach including PPSNet (lines 4-7)

Require: I, p, d, u, K
 1: $RGB_{feats} = DINOv2\_Encoder(I)$
 2: $D_{init} = DPT\_Decoder(RGB_{feats})$
 3: $PPS_{feats} = CalculatePPL(D_{init}, p, d, u, K)$ ▷ involves eqn. 1
 4: $x_{combo} = CrossAttention(RGB_{feats}, PPS_{feats})$ ▷ involves eqn. 5
 5: $D_{mod} = FiLM\_Mod(D_{init}, x_{combo})$ ▷ involves eqn. 6
 6: $\Delta_D = UNet(D_{mod})$
 7: $D_{ref} = D_{init} + \Delta_D$

3.4 Teacher-Student Transfer Learning for Sim2Real

Teacher network. We first train our proposed network architecture that combines a foundational monocular depth estimation network followed by PPSNet, as shown in algorithm 1 on synthetic C3VD [8] data with supervision. Specifically, we train with scale-shift invariant ($\mathcal{L}_{SSI}$), depth regularization ($\mathcal{L}_{REG}$), and virtual-normal ($\mathcal{L}_{VNL}$) losses described in prior works [10, 51], along with our proposed $\mathcal{L}_{PPS-sup}$ loss. While this network produces state-of-the-art performance on the C3VD dataset we observe that it often produces errors and artifacts on real clinical data. This is a common problem for any neural network trained on synthetic data and tested on real data due to the Sim2Real gap.

To address this problem we propose Teacher-Student transfer learning, where we train a student network on real data with self-supervision and the guidance of the Teacher network. A similar approach is also utilized by a recent state-of-the-art monocular depth estimation algorithm [67] that leverages large-scale data (1.5 million labeled and 62 million unlabeled images) and trains using a self-supervised loss approach with semantic losses. Such a large-scale unlabeled data is unavailable for endoscopy and semantic losses are meaningless for colons. Instead, we propose a teacher-student transfer learning that utilizes illumination information to train on ∼18,000 real clinical images [2,40] using self-supervision.
Student network. Our Student network has the same architecture as the Teacher network. The Student model is trained using both labeled synthetic

data ($\mathcal{D}_L$) and unlabeled clinical data ($\mathcal{D}_U$) in an interleaved manner, which provides more stability during training. When using labeled synthetic data we use the same supervised loss functions as the teacher. However, when training on unlabeled real data we have no supervision w.r.t. ground-truth depth and normal map. Instead, we use the proposed self-supervised loss function $\mathcal{L}_{PPS-corr}$ for training. However, self-supervision on its own is not powerful enough to produce accurate depth estimates due to inherent ambiguity in training. Thus we use guidance from the Teacher network in terms of pseudo-supervision, where the output of the Teacher network for an input clinical image is used as a pseudo ground-truth for computing the supervised loss functions over depth and PPS. Experimental analysis in Figure 5 shows how teacher-student transfer learning improves depth estimation quality on real clinical data. Further details regarding our teacher-student learning approach, including a corresponding algorithm table, can be found in our supplementary materials.

3.5   Implementation details

We implement our approach in PyTorch [47] and use the AdamW [27,38] optimizer coupled with the OneCycleLR policy [62] for training. For the OneCycleLR policy, we choose a maximum learning rate of 1e-5 and train for 20 epochs with a batch size of 8 on 4 x NVIDIA RTX A6000s (48GB each). We also utilize the initial weights from [67] as a starting point for all of our baselines and results in order to leverage the rich, prior knowledge of state-of-the-art monocular depth estimation methods. It takes approximately 5 hours to train our teacher model and approximately 10 hours to train our student model. The difference in training times is because the student model is exposed to significantly more data as a result of being trained on unlabeled clinical data. At training time the DINOv2 encoder, DPT decoder, and the refinement module are all trainable. Further details regarding our utilized losses, including loss weights, can be found in our supplementary materials.

4   Results

4.1   Experimental Setup

Datasets. We train and test on separate splits of the C3VD [8] dataset and our clinical dataset composed of oblique and en face views [2] and Colon10K [40]. C3VD includes 22 video sequences amounting to 10,015 frames with comprehensive ground truth annotations. Our clinical dataset includes 80 sequences with oblique views (7,293 frames), 14 sequences with en-face views (832 frames), and 20 sequences with down-the-barrel (axial) views (10,216 frames). For training and testing with C3VD, we re-use the splits noted in the supplementary materials of LightDepth [54]. Further details are provided in our supplementary.

Baselines. We compare our approach to the state-of-the-art monocular depth estimation techniques on endoscopy videos. • UNet [57] is a meaningful baseline when provided with ground truth depth due to its ability to capture multi-scale contextual information and precisely localize features through its symmetric expanding path. • MonoDepth2 [15] and NormDepth [64] are promising,

Table 1: We compare our approach with existing monocular depth estimation techniques developed for endoscopy videos by testing on the synthetic C3VD dataset. Ours-Student trains using both C3VD and real data. Best results are shown in bold. Second best results are underlined.

Architecture	Supervision	$RMSE \downarrow$	$RMSE_{log} \downarrow$	$AbsRel \downarrow$	$SqRel \downarrow$	$\delta < 1.1 \uparrow$
UNet [57]	GT	6.71	0.1628	0.126	1.72	0.63
MonoDepth2 [15]	SSL	9.11	0.221	0.192	2.64	0.43
NormDepth [64]	SSL	9.72	0.251	0.164	2.05	0.42
Wang et al. [64]	GT & SSL	7.51	0.188	0.155	1.53	0.48
LightDepth DPT [54]	SSL	6.55	0.202	0.0780	1.81	0.56
DepthAnything [67]	GT & SSL	12.76	0.318	0.283	6.41	0.34
Ours - Backbone	GT	2.79	0.0785	0.0622	0.238	0.82
Ours - Teacher	GT	<u>2.15</u>	<u>0.0705</u>	<u>0.0529</u>	<u>0.147</u>	<u>0.87</u>
Ours - Student	GT & SSL	**2.06**	**0.0656**	**0.0491**	**0.140**	**0.89**

self-supervised approaches that leverage multi-view photometric consistency to enforce geometric constraints across different viewpoints. • We also compare with *Wang et al.* [64] which includes NormDepth as a part of a SLAM pipeline for joint normal and depth refinement. • LightDepth [54] is a recent approach that is completely self-supervised and leverages the principle of illumination decline alongside a photometric model. • We also compare to the current state-of-the-art in monocular depth estimation, Depth Anything [67], which is not trained on endoscopy images. • We then present Ours-Backbone, the backbone model that fine-tunes a pre-trained DINOv2 [44] encoder and DPT [50] decoder used in [67] using supervised loss functions on the depth map, without any PPS based losses. • We then present Ours-Teacher network, which combines Ours-Backbone with the PPSNet depth refinement module and trains on C3VD with supervised depth and PPS losses. • Finally, Ours-Student model, has the same architecture as Ours-Teacher but is trained on both C3VD and real clinical data using supervised and self-supervised PPS loss functions respectively.

Metrics. We evaluate commonly used depth metrics such as root-mean-squared error ($RMSE$), log-scale RMSE ($RMSE_{log}$), absolute relative error ($AbsRel$), squared relative error ($SqRel$) multiplied by a factor of 1000, and $\delta < 1.1$ which indicates percentage of pixels within 10% of the actual depth values.

4.2   Comparison with Existing Approaches

Quantitative Evaluation. In Table 1, we compare to the state-of-the-art approaches on the C3VD [8] dataset that has paired ground truth depth. We include both Ours-Teacher model (trained only on C3VD data) and Ours-Student model (trained on both C3VD data and clinical data) for comparison. Unless indicated otherwise, all other baselines were trained only on C3VD data.
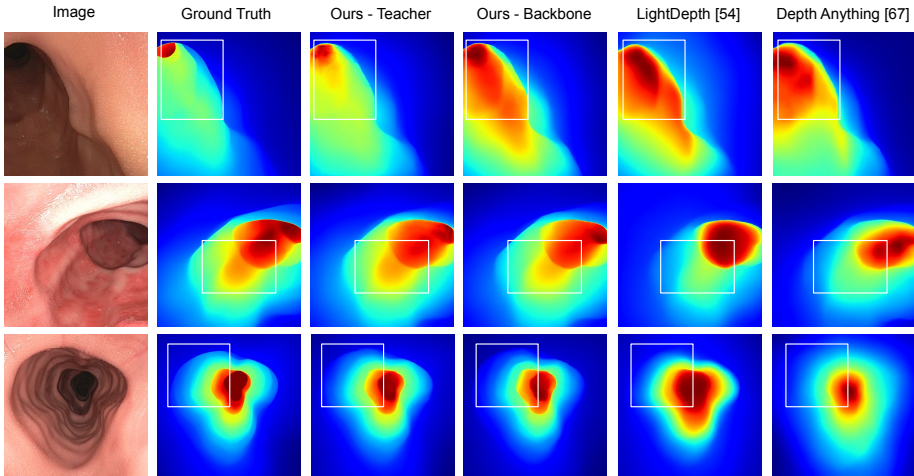
Fig. 4: Qualitative evaluation on the C3VD dataset. Red = further distance from the camera and blue is closer. Regions to note during visual comparison are outlined in white. Ours-Teacher performs best, improving upon Ours-Bakcbone due to our proposed depth refinement PPSNet and self-supervised PPS loss.

Qualitative Results. We show qualitative results in the form of relative depth maps for both the C3VD dataset and our clinical dataset composed of down-the-barrel (axial), oblique, and en-face views. As shown in Figs. 4 and 5, our method produces higher-quality depth maps with fewer discrepancies in the case of C3VD and clinical data.

Ablation Studies. Role of PPSNet and $\mathcal{L}_{PPS-sup}$ in improving Ours-Teacher over Ours-Backbone.

In Table 1 and Figure 4 and 5 we showed that Ours-Teacher is better than Ours-Backbone. We propose two distinct modifications over Ours-Backbone, involving depth refinement module PPSNet and supervised loss function $\mathcal{L}_{PPS-sup}$. In Table 2 we show that both PPSNet and $\mathcal{L}_{PPS-sup}$ help in improving the performance of Ours-Teacher (row 4) over Ours-Backbone (row 1).

Table 2: We show that both PPSNet depth refinement and $\mathcal{L}_{PPS-sup}$ loss helps in improving the performance of Ours-Teacher (row 4) over Ours-Backbone (row 1).

$\mathcal{L}_{PPS-sup}$	PPSNet	RMSE ↓	AbsRel ↓
✗	✗	2.79	0.0621
✓	✗	2.21	0.0518
✗	✓	2.60	0.0617
✓	✓	2.15	0.0529

Role of features extracted from PPS representation in the depth refinement network PPSNet. In Table 3 we show that modulating initial depth features extracted from input images using DINOv2 with RGB features (RGB_{feats}) and PPS features (PPS_{feats}) improves the depth refinement performance.

Role of teacher-student learning. In Table 1 and Figure 5 we showed that Ours-Student improves over Ours-Teacher, noticeably on real clinical data since

Ours-Teacher is only trained on C3VD and hence fails to generalize well on real clinical data. The strong performance of Ours-Student can be attributed to two key contributions. Firstly we show in Table 4 that using self-supervised PPS loss function $\mathcal{L}_{PPS-corr}$ on real data helps in improving the performance, even when testing on synthetic data.

Additionally, we observe that having a strong teacher module trained on synthetic data provides a general prior for colon geometry and has a measurable benefit over just using DINOv2 features from a teacher network. Finally, we note that the most optimal performance across all metrics is obtained by using both our proposed teacher and $\mathcal{L}_{PPS-corr}$.

Table 3: We show that RGB_{feats} and PPS_{feats} improves depth refinement on C3VD when used for modulation.

RGB_{feats}	PPS_{feats}	RMSE ↓	AbsRel ↓
✓	✗	2.29	0.0531
✗	✓	2.24	0.0551
✓	✓	2.15	0.0529

Table 4: We show that both self-supervised $\mathcal{L}_{PPS-corr}$ and a strong teacher network Ours-Teacher helps in improving the performance of Ours-Student.

Teacher	$\mathcal{L}_{PPS-corr}$	RMSE ↓	AbsRel ↓
DINOv2	✗	2.55	0.0625
DINOv2	✓	2.31	0.0533
Ours	✗	2.28	0.0544
Ours	✓	2.06	0.0491

5   Limitations and Discussion

Despite the success of our approach and its ability to be extended to any monocular depth estimation method through the adoption of our learning objectives and our refinement module, we note that the formulation of per-pixel lighting that we utilize does not adequately model albedo and does not directly address in-view, specular lighting effects that can lead to misinterpreted lighting cues and ultimately inaccurate depth estimations. To remedy this, we opt to detect and mask out specular lighting effects. Future works should aim to model such effects and evaluate their impact on downstream tasks such as 3D mesh reconstruction. We plan to utilize our approach to depth estimation in a future pipeline that integrates tracking and depth prediction as a part of a SLAM system.

6   Conclusion

This paper presents a novel approach to leveraging per-pixel lighting (PPL) information toward enhanced depth estimation in colonoscopy. We use supervised and self-supervised variants of loss functions that use an analytically computed per-pixel shading (PPS) field, which is also used as an input into our refinement network that conditions an initial predicted depth on a combination of RGB features and PPS features. Our approach achieves state-of-the-art results on the C3VD dataset while enabling the estimation of high-quality depth maps on real clinical data. We do not feel that our approach is limited to colonoscopy, nor bio-applications such as endoscopy for that matter. Our approach can be useful in other bio-applications such as bronchoscopy, which we demonstrate in our supplementary materials, as well as more mainstream contexts for depth estimation (e.g., in egocentric vision).
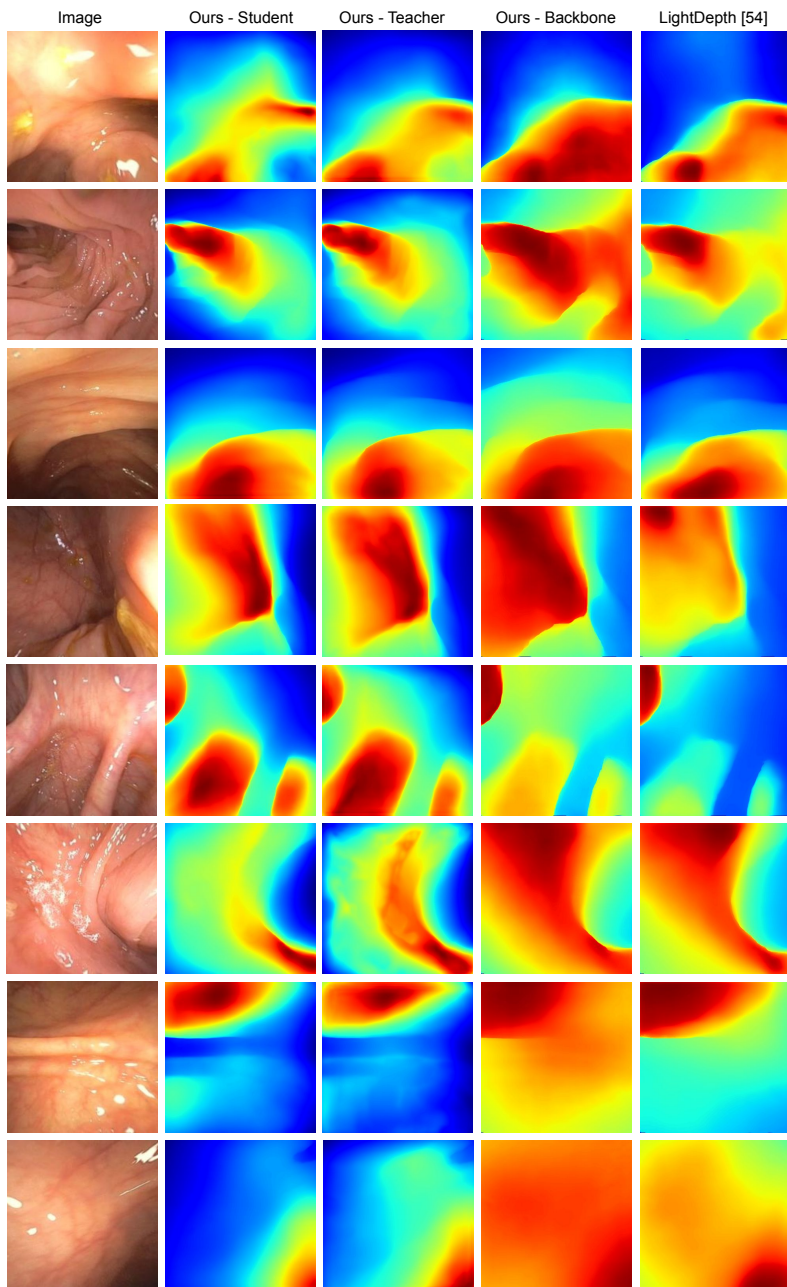
Fig. 5: Qualitative evaluation on clinical data. Red = further distance from the camera and blue is closer.

A   Overview of Appendices

Our appendices contain the following additional details and results:

- In Sec. B, we present point cloud results from our approach. We also provide point cloud results from LightDepth [54] for comparison.
- In Sec. C, we provide additional information regarding our approach. Sec. C.1 includes additional details about the proposed per-pixel lighting (PPL) representation and its usage in our proposed approach. In Sec. C.2, we provide additional details regarding our approach to teacher-student learning, including a full algorithm table. Sec. C.3 contains additional details regarding our implementation, including our chosen loss weights and backbone model size.
- Sec. D includes additional information regarding our experiments, chiefly the training splits we utilized that match [54].
- Sec. E describes our limited code release included as a part of these supplementary materials. The code includes our proposed loss functions and model files for reference, as well as pre-trained models. We will release our full code, including the training code and various baselines, in the near future.
- Sec. F includes additional information regarding our implementation of Light-Depth [54].
- We provide additional, qualitative results on bronchoscopy data in Sec. G.
- Sec. H includes additional details regarding the clinical dataset utilized in our work and to be released in the near future.

B   Point Clouds

We present point cloud results from our monocular depth estimation approach in order to show that our produced depths can eventually be used for the task of 3D mesh reconstruction and subsequent 3D analysis. In Fig. 6 it can be seen that our produced point clouds are superior to the point clouds produced by our re-implementation of LightDepth [54]. We also provide the raw *.ply* files for each point cloud in Fig. 6 as a part of our supplementary materials.
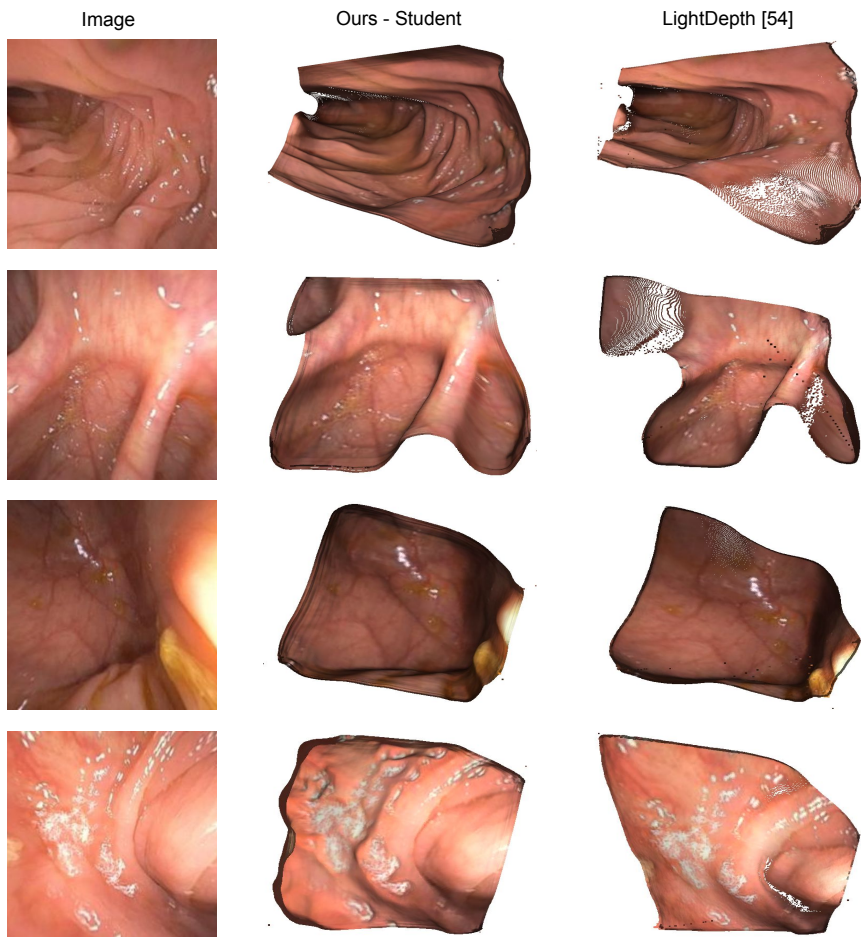

Fig. 6: A qualitative view of 3D point clouds generated by our approach and in contrast to LightDepth [54]. We also provide the raw *.ply* files for each point cloud as a part of our supplementary materials.

C   Additional Info on Methods

C.1   Per-Pixel Lighting Representation and Losses

In our main paper, we noted that we utilize the conventional pinhole camera framework, positioned at the origin in the global coordinate system and looking down the z-axis. This camera is defined by the intrinsic matrix K. A point in the world space, denoted as $X = (x, y, z)$, is mapped to a pixel coordinate (u, v) using:

$$(u, v, 1)^T \sim K(x, y, z)^T . \tag{7}$$

Subsequently, we noted that we only take into account visible surfaces of the colon, and therefore assume that the depth map and corresponding normal map is a complete description of in-view surfaces. $X(u, v) \in \mathbb{R}^3$ describes the 3D location of a pixel (u, v) on the surface of the organ. The depth map can then be defined by $D(u, v) = X(u, v)_3$, corresponding to the z-component of $X(u, v)$. $X(u, v)$ itself can be recovered from the following:

$$X(u, v) = D(u, v)K^{-1}(u, v, 1)^T . \tag{8}$$

For completeness, we now also note that given $n(X)$ is the normal at the point X, then the normal map can be defined by $N(u, v) = n(X(u, v))$. N is subsequently computed based on the following:

$$N = \frac{\frac{\partial X}{\partial u} \times \frac{\partial X}{\partial v}}{\left\| \frac{\partial X}{\partial u} \times \frac{\partial X}{\partial v} \right\|} . \tag{9}$$

Additionally, we note that our learning objective $\mathcal{L}_{PPS-corr}$ can be formulated as follows:

$$\mathcal{L}_{PPS-corr} = 1 - \frac{\sum_{h=1}^{H} \sum_{w=1}^{W} (I_{gray_{hw}} - \bar{I})(PPL_{F_{hw}} - \overline{PPL}_F)}{\sqrt{\sum_{h=1}^{H} \sum_{w=1}^{W} (I_{gray_{hw}} - \bar{I})^2 \sum_{hw} (PPL_{F_{hw}} - \overline{PPL}_F)^2}} \tag{10}$$

The simplified version of this formulation is presented in our main paper as follows:

$$\mathcal{L}_{PPS-corr} = 1 - \text{corr}(I_{gray}, PPL_F) \tag{11}$$

As noted in the main paper, the self-supervised loss function variant will enable us to train on real clinical data where ground-truth depth information is unavailable. Code implementations of both our supervised and self-supervised loss function variants are included as a part of these supplementary materials.

C.2   Teacher-Student Transfer Learning for Sim2Real

Our approach with teacher-student training on labeled data $\mathcal{D}_L$ and unlabeled data $\mathcal{D}_U$ can be summarized in algo. 2 as follows:

Where the losses used alongside our proposed losses $\mathcal{L}_{PPS-sup}$ and $\mathcal{L}_{PPS-corr}$ correspond to the scale-shift invariant ($\mathcal{L}_{SSI}$), regularization ($\mathcal{L}_{REG}$), and virtual-normal ($\mathcal{L}_{VNL}$) losses described in prior works [10, 51].

Algorithm 2 Teacher-Student Learning for Endoscopic Data

Require: labeled synthetic dataset: $\mathcal{D}_L$, unlabeled real dataset: $\mathcal{D}_U$
1: Initialize teacher and student models with identical architectures:
      M_T (Teacher), M_S (Student) $\leftarrow InitializeModels()$
2: Train M_T on $\mathcal{D}_L$: $M_T \leftarrow Train(M_T, \mathcal{D}_L)$ $\triangleright$ Note: Refers to algo. 1.
3:    $\mathcal{L} = \alpha_{SSI}\mathcal{L}_{SSI} + \alpha_{REG}\mathcal{L}_{REG} + \alpha_{VNL}\mathcal{L}_{VNL} + \alpha_{PPS-sup}\mathcal{L}_{PPS-sup}$
4: Freeze M_T to prevent further updates
5: Prepare mixed dataset $\mathcal{D}_M$ combining $\mathcal{D}_L$ and $\mathcal{D}_U$.
6: **for** each batch b in $\mathcal{D}_M$ **do**
7:    **if** b is from $\mathcal{D}_L$ **then**
8:       Train M_S on $\mathcal{D}_L$: $M_S \leftarrow Train(M_S, \mathcal{D}_L)$
9:         $\mathcal{L} = \alpha_{SSI}\mathcal{L}_{SSI} + \alpha_{REG}\mathcal{L}_{REG} + \alpha_{VNL}\mathcal{L}_{VNL} + \alpha_{PPS-sup}\mathcal{L}_{PPS-sup}$
10:   **else if** b is from $\mathcal{D}_U$ **then**
11:      Train M_S on $\mathcal{D}_U$: $M_S \leftarrow Train(M_S, \mathcal{D}_U)$
12:        $\mathcal{L} = \alpha_{SSI}\mathcal{L}_{SSI*} + \alpha_{REG}\mathcal{L}_{REG*} + \alpha_{VNL}\mathcal{L}_{VNL*} + \alpha_{PPS-corr}\mathcal{L}_{PPS-corr}$
13:        Compute $\mathcal{L}_{SSI*}$, $\mathcal{L}_{REG*}$, and $\mathcal{L}_{VNL*}$ using pseudo-supervision from M_T.
14:   **end if**
15: **end for**

C.3   Implementation

As a part of training our approach, we use the following α values that describe loss weights as shown in algo. 2: $\alpha_{SSI} = 1.0$, $\alpha_{REG} = 0.1$, $\alpha_{VNL} = 10.0$, $\alpha_{PPS-sup} = 1.0$, and $\alpha_{PPS-corr} = 1.0$. We utilize the ViT-Small version of the backbone architecture that's also utilized in Depth Anything [67].

D   Experiments

We utilized the same dataset splits as LightDepth [54] and include the exact sequences used for the C3VD [8] dataset in Tab. 5. Our clinical dataset includes 80 sequences with oblique views (7,293 frames), 14 sequences with en-face views (832 frames), and 20 sequences with down-the-barrel (axial) views (10,216 frames). The oblique views and en-face views are from a to-be-released dataset described in Sec. H. The down-the-barrel (axial) views are from the Colon10K [40] dataset. The exact clinical sequences used for training and testing are included as *train.txt* and *val.txt* files in our supplementary materials folder.

E   Limited Code Release

Our limited code release can be accessed via a link to a GitHub repo on our project website: `https://ppsnet.github.io/`. The GitHub repo contains a *PPSNet.py* file that serves as the full model file for our proposed approach. *calculate_PPL.py* from [32] is also included for reference alongside our proposed loss functions in *PPS_losses.py*. We also provide limited evaluation code for

Table 5: Dataset Splits for C3VD [8]

Sequence	Texture	Video	Frames	Set
Cecum	1	b	765	Train
Cecum	2	b	1120	Train
Cecum	2	c	595	Train
Cecum	4	a	465	Train
Cecum	4	b	425	Train
Sigmoid Colon	1	a	800	Train
Sigmoid Colon	2	a	513	Train
Sigmoid Colon	3	b	536	Train
Transcending	1	a	61	Train
Transcending	1	b	700	Train
Transcending	2	b	102	Train
Transcending	2	c	235	Train
Transcending	3	b	214	Train
Transcending	4	b	595	Train
Descending Down	4	a	74	Train
Cecum	1	a	276	Test
Cecum	2	a	370	Test
Cecum	3	a	730	Test
Sigmoid	3	a	610	Test
Transcending	2	a	194	Test
Transcending	3	a	250	Test
Transcending	4	a	384	Test
Descending Up	4	a	74	Test

the C3VD [8] dataset and pre-trained models. We will release our full codes, including the training code and various baselines, in the near future.

F   LightDepth Implementation

For the purpose of comparison to state-of-the-art monocular depth estimation specific to endoscopy, we attempted to implement LightDepth [54] on our own, as the authors did not release their code. The authors describe an architecture with two ViT branches: one initialized with the weights from DPT-Hybrid, used for depth prediction; and one trained from scratch, used for prediction of albedo. Specifically, this albedo predictor computes hue and saturation values for each pixel, and then converts to RGB color space from HSV assuming V=100%. We were unable to successfully implement this two-headed depth and albedo prediction. Instead, we compute the albedo ρ by analytically solving the per-

pixel rendering equation used (eq. 3 in [54]):

$$\mathcal{I}(d_i, \rho_i, g) = \left(\frac{\sigma_0}{||d_i \mathbf{r}_i - \mathbf{x}_l||^2} R(\psi_i) \cos \theta_i \rho_i g \right)^{1/\gamma}$$

with ψ_i, $R(\psi_i)$, and θ_i depending only on constants and the predicted depth map (whose value at pixel i is d_i).

Assuming a colocated light and camera and setting $\sigma_0 = g = 1$, we can rearrange to get:

$$\rho_i = \frac{d_i^2 \mathcal{I}_i^\gamma}{R(\psi_i) \cos \theta_i}$$

where $\mathcal{I}_i$ is the R, G, or B value of the ith pixel, and we apply this separately for each channel. This calculated ρ_i replaced the predicted albedo ρ_i in all loss calculations. We additionally add a regularization loss

$$\mathcal{L}_{albvar} = \frac{1}{3} (\mathrm{Var}(\rho^R) + \mathrm{Var}(\rho^G) + \mathrm{Var}(\rho^B))$$

where $\mathrm{Var}(\rho^R)$ is the variance of the R channel calculated albedo value over the entire batch, and likewise for G and B. This provided a slight improvement in results, the idea being that in endoscopy the surface albedo should be fairly uniform.

Starting from the pretrained DPT-Hybrid weights from [10], we fine-tuned our LightDepth implementation for a single epoch on the train split of C3VD used in [54] at a learning rate of 4e-6, as any further training was found to overfit and reduce accuracy.

G   Qualitative Results on Bronchoscopy

In Fig. 7, we present qualitative depth map results on bronchoscopy data. Despite not on bronchoscopy data, we note that 'Ours - Student' and 'Ours - Backbone' both provide superior results to the baseline provided by Depth Anything [67], with 'Ours - Student' being slightly better than 'Ours - Backbone' in certain areas with relatively farther depths. We plan to explore bronchoscopy data as an additional bio-application in subsequent work. We will release a small dataset of bronchsocopy frames for qualitative evaluation upon this work's acceptance.
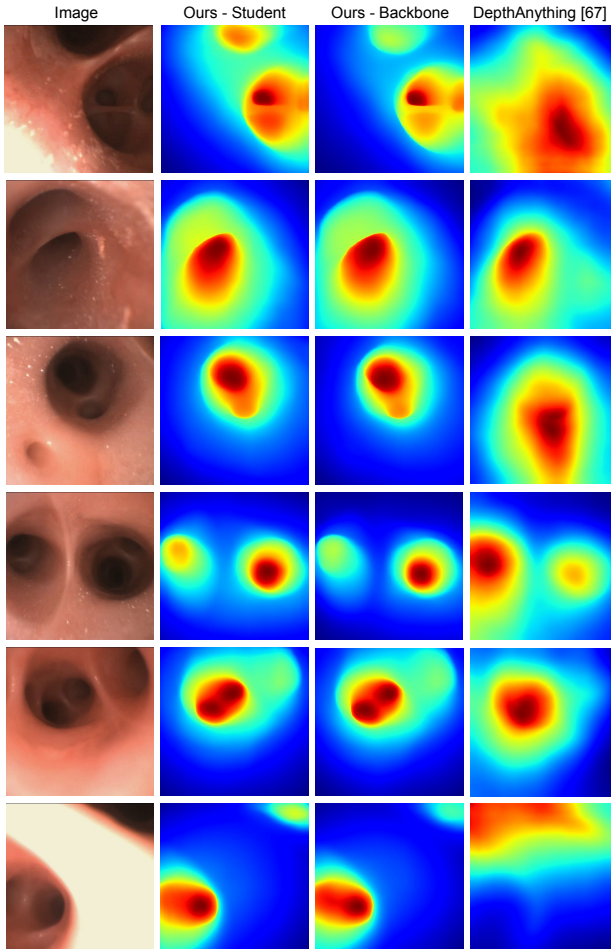
Fig. 7: Qualitative evaluation on bronchoscopy data. Red = further distance from the camera and blue is closer. Note that 'Ours - Student' and 'Ours - Backbone' provide significantly higher quality depth estimations than Depth Anything [67].

H Clinical Dataset

The clinical dataset used in our work consists of 80 sequences with oblique views (7,293 frames), 14 sequences with en-face views (832 frames), and 20 sequences with down-the-barrel (axial) views (10,216 frames). The down-the-barrel (axial) views are from the Colon10K [40] dataset. The oblique views and en-face views are from a dataset which will be released in a separate submission [2] to a different conference. The anonymized submission in question is included as a part of our supplementary materials.

References

1. Andolfi, M., Potenza, R., Capozzi, R., Liparulo, V., Puma, F., Yasufuku, K.: The role of bronchoscopy in the diagnosis of early lung cancer: a review. Journal of thoracic disease **8**(11), 3329 (2016)

2. Anonymous: Structure-preserving image translation for depth estimation in colonoscopy (2024), paper submitted to another conference that releases a portion of the clinical data used in this paper.

3. Azagra, P., Sostres, C., Ferrandez, A., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gomez-Rodriguez, J.J., Elvira, R., Lopez, J., Oriol, C., Civera, J., Tardos, J.D., Murillo, A.C., Lanas, A., Montiel, J.M.M.: Endomapper dataset of complete calibrated endoscopy procedures. Scientific Data **10**(1) (Oct 2023). `https://doi.org/10.1038/s41597-023-02564-7`, `http://dx.doi.org/10.1038/s41597-023-02564-7`

4. Bae, G., Budvytis, I., Cipolla, R.: Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. arXiv preprint arXiv:2210.03676 (2022)

5. Bae, G., Budvytis, I., Yeung, C.K., Cipolla, R.: Deep multi-view stereo for dense 3d reconstruction from monocular endoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 774–783. Springer (2020)

6. Basri, R., Jacobs, D., Kemelmacher, I.: Photometric stereo with general, unknown lighting. International Journal of computer vision **72**, 239–257 (2007)

7. Berci, G., Forde, K.A.: History of endoscopy. Surgical endoscopy **14**(1), 5–15 (2000)

8. Bobrow, T.L., Golhar, M., Vijayan, R., Akshintala, V.S., Garcia, J.R., Durr, N.J.: Colonoscopy 3d video dataset with paired depth from 2d-3d registration. Medical Image Analysis **90**, 102956 (Dec 2023). `https://doi.org/10.1016/j.media.2023.102956`, `http://dx.doi.org/10.1016/j.media.2023.102956`

9. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. Advances in neural information processing systems **29** (2016)

10. Eftekhar, A., Sax, A., Bachmann, R., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans (2021)

11. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)

12. Faro, A., Wood, R.E., Schechter, M.S., Leong, A.B., Wittkugel, E., Abode, K., Chmiel, J.F., Daines, C., Davis, S., Eber, E., et al.: Official american thoracic society technical standards: flexible airway endoscopy in children. American journal of respiratory and critical care medicine **191**(9), 1066–1080 (2015)

13. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)

14. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International Conference on Machine Learning. pp. 1607–1616. PMLR (2018)

15. Godard, C., Aodha, O.M., Firman, M., Brostow, G.: Digging into self-supervised monocular depth estimation (2019)

16. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**, 1789–1819 (2021)

17. Grand View Research: Endoscopy procedures estimates market volume, share & trends analysis report by application (gastroscopy, colonoscopy, laparoscopy), by region (north america, asia pacific, europe), and segment forecasts, 2023 - 2030. https://www.grandviewresearch.com/industry-analysis/endoscopy-procedures-estimates-market-report (2023)

18. Heimann, T., Meinzer, H.P.: Statistical shape models for 3d medical image segmentation: a review. Medical image analysis **13**(4), 543–563 (2009)

19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

20. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision **75**, 151–172 (2007)

21. Hwang, S.J., Park, S.J., Kim, G.M., Baek, J.H.: Unsupervised monocular depth estimation for colonoscope system using feedback network. Sensors **21**(8), 2691 (2021)

22. Kaminski, M.F., Thomas-Gibson, S., Bugajski, M., Bretthauer, M., Rees, C.J., Dekker, E., Hoff, G., Jover, R., Suchanek, S., Ferlitsch, M., et al.: Performance measures for lower gastrointestinal endoscopy: a european society of gastrointestinal endoscopy (esge) quality improvement initiative. United European gastroenterology journal **5**(3), 309–334 (2017)

23. Kaufman, A., Wang, J.: 3d surface reconstruction from endoscopic videos. In: Visualization in Medicine and Life Sciences, pp. 61–74. Springer (2008)

24. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation (2023)

25. Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 573–590 (2018)

26. Kim, S.Y., Zhang, J., Niklaus, S., Fan, Y., Chen, S., Lin, Z., Kim, M.: Layered depth refinement with mask guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3855–3865 (2022)

27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

28. Kuipers, E.J., Rösch, T., Bretthauer, M.: Colorectal cancer screening—optimizing current strategies and new directions. Nature reviews Clinical oncology **10**(3), 130–142 (2013)

29. Kusupati, U., Cheng, S., Chen, R., Su, H.: Normal assisted stereo depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2189–2199 (2020)

30. Kwon, H., Tai, Y.W., Lin, S.: Data-driven depth map refinement via multi-scale sparse representation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 159–167 (2015)

31. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)

32. Lichy, D., Sengupta, S., Jacobs, D.W.: Fast light-weight near-field photometric stereo (2022)

33. Lichy, D., Wu, J., Sengupta, S., Jacobs, D.W.: Shape and material capture at home. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6123–6133 (2021)

34. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: Sift flow: Dense correspondence across different scenes. In: Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10. pp. 28–42. Springer (2008)
35. Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE transactions on medical imaging **39**(5), 1438–1447 (2019)
36. Logothetis, F., Budvytis, I., Mecca, R., Cipolla, R.: A cnn based approach for the near-field photometric stereo problem. arXiv preprint arXiv:2009.05792 (2020)
37. Logothetis, F., Mecca, R., Cipolla, R.: A differential volumetric approach to multiview photometric stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1052–1061 (2019)
38. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
39. Luo, H., Hu, Q., Jia, F.: Details preserved unsupervised depth estimation by fusing traditional stereo knowledge from laparoscopic images. Healthcare technology letters **6**(6), 154–158 (2019)
40. Ma, R., McGill, S.K., Wang, R., Rosenman, J., Frahm, J.M., Zhang, Y., Pizer, S.: Colon10k: a benchmark for place recognition in colonoscopy. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1279–1283. IEEE (2021)
41. Ma, R., Wang, R., Zhang, Y., Pizer, S., McGill, S.K., Rosenman, J., Frahm, J.M.: Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy. Medical image analysis **72**, 102100 (2021)
42. Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J., Montiel, J.M.M.: Orbslam-based endoscope tracking and 3d reconstruction. In: Computer-Assisted and Robotic Endoscopy: Third International Workshop, CARE 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 3. pp. 72–83. Springer (2017)
43. Nierengarten, M.B.: Colonoscopy remains the gold standard for screening despite recent tarnish: Although a recent study seemed to indicate that colonoscopies are not as effective as once thought at detecting colorectal cancer, a closer look at the study clears the confusion (2023)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024)
45. 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**, 102058 (2021)
46. Papadhimitri, T., Favaro, P.: Uncalibrated near-light photometric stereo (2014)
47. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
48. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer (2017)

49. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 283–291 (2018)
50. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (October 2021)
51. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer (2020)
52. Rau, A., Bano, S., Jin, Y., Azagra, P., Morlana, J., Sanderson, E., Matuszewski, B.J., Lee, J.Y., Lee, D.J., Posner, E., et al.: Simcol3d–3d reconstruction during colonoscopy challenge. arXiv preprint arXiv:2307.11261 (2023)
53. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J., Civera, J.: Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE Robotics and Automation Letters **6**(4), 7225–7232 (2021)
54. Rodriguez-Puigvert, J., Batlle, V.M., Montiel, J.M.M., Martinez-Cantin, R., Fua, P., Tardos, J.D., Civera, J.: Lightdepth: Single-view depth self-supervision from illumination decline (2023)
55. Rodriguez-Puigvert, J., Recasens, D., Civera, J., Martinez-Cantin, R.: On the uncertain single-view depths in colonoscopies. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 130–140. Springer (2022)
56. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
57. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
58. Rossi, M., Gheche, M.E., Kuhn, A., Frossard, P.: Joint graph-based depth refinement and normal estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12154–12163 (2020)
59. Santo, H., Waechter, M., Matsushita, Y.: Deep near-light photometric stereo for spatially varying reflectances. In: European Conference on Computer Vision. pp. 137–152. Springer (2020)
60. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence **31**(5), 824–840 (2008)
61. Scarzanella, M.V., Sugiura, T., Kaneko, T., ichiro Koto, S.: Deep monocular 3d reconstruction for assisted navigation in bronchoscopy. International Journal of Computer Assisted Radiology and Surgery **12**, 1089–1099 (2017), `https://api.semanticscholar.org/CorpusID:26313018`
62. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multidomain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
63. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
64. Wang, S., Zhang, Y., McGill, S.K., Rosenman, J.G., Frahm, J.M., Sengupta, S., Pizer, S.M.: A surface-normal based neural framework for colonoscopy reconstruction (2023)

65. Widya, A.R., Monno, Y., Imahori, K., Okutomi, M., Suzuki, S., Gotoda, T., Miki, K.: 3d reconstruction of whole stomach from endoscope video using structure-from-motion. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 3900–3904. IEEE (2019)
66. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5354–5362 (2017)
67. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data (2024)
68. Yang, W., Chen, G., Chen, C., Chen, Z., Wong, K.Y.K.: Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In: European Conference on Computer Vision. pp. 266–284. Springer (2022)
69. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
70. Zhang, Y., Frahm, J.M., Ehrenstein, S., McGill, S.K., Rosenman, J.G., Wang, S., Pizer, S.M.: Colde: A depth estimation framework for colonoscopy reconstruction (2021)
71. Zhao, D., Lichy, D., Perrin, P.N., Frahm, J.M., Sengupta, S.: Mvpsnet: Fast generalizable multi-view photometric stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12525–12536 (2023)
72. Zhao, Q., Price, T., Pizer, S., Niethammer, M., Alterovitz, R., Rosenman, J.: The endoscopogram: A 3d model reconstructed from endoscopic video frames. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I 19. pp. 439–447. Springer (2016)