LightDepth: Single-View Depth Self-Supervision from Illumination Decline

Javier Rodríguez-Puigvert[1*] Víctor M. Batlle[1*] J.M.M. Montiel[1] Ruben Martinez-Cantin[1]
Pascal Fua[2] Juan D. Tardós[1] Javier Civera[1]

[1]I3A - Universidad de Zaragoza [2]École Polytechnique Fédérale de Lausanne

Abstract

Single-view depth estimation can be remarkably effective if there is enough ground-truth depth data for supervised training. However, there are scenarios, especially in medicine in the case of endoscopies, where such data cannot be obtained. In such cases, multi-view self-supervision and synthetic-to-real transfer serve as alternative approaches, however, with a considerable performance reduction in comparison to supervised case. Instead, we propose a single-view self-supervised method that achieves a performance similar to the supervised case. In some medical devices, such as endoscopes, the camera and light sources are co-located at a small distance from the target surfaces. Thus, we can exploit that, for any given albedo and surface orientation, pixel brightness is inversely proportional to the square of the distance to the surface, providing a strong single-view self-supervisory signal. In our experiments, our self-supervised models deliver accuracies comparable to those of fully supervised ones, while being applicable without depth ground-truth data.

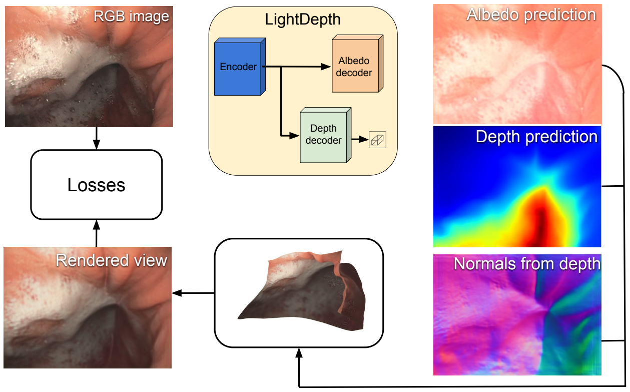

Figure 1. **Single-view depth self-supervision in LightDepth.** A two-headed deep network predicts albedo and depth from a single image and estimates surface normals from predicted depths. These are used to render a new image, that takes into account illumination decline and the endoscope's photometric calibration, and can be compared to the original one. Minimizing the difference between the original and rendered images is used at training time to compute the network weights and at inference time to refine the depth predictions.

1. Introduction

Minimally invasive medical procedures such as gastroscopies, colonoscopies and bronchoscopies rely on endoscopes that should be as small as possible. As a result, they usually house a single camera and several light points, but neither depth nor stereo cameras. 3D reconstruction is relevant in endoscopies, as it may unlock several functionalities such as the accurate estimation of the size and shape of tumors. However, both single- and multi-view depth estimation methods present significant challenges in this domain. The lack of sufficient depth annotated data hinders the use of supervised depth learning. The presence of fluids that either obscure the view or generate specularities, the sudden illumination changes, the paucity of texture and the surface deformations hamper multi-view methods both for self-supervising deep networks and for geometry estimation. Real in-body textures and fluids are hard to simulate realistically, and the synthetic-to-real gap may be large.

In this work, we propose a novel approach to depth in endoscopies that overcomes all the above challenges related to depth supervision, multi-view estimation and synthetic-to-real gaps. Our key insight is that, by exploiting a key property of endoscopic imagery, we can provide strong depth self-supervision signals from just one view. In endoscopes, the light source is rigidly located next to the camera and is close to the surface to be reconstructed. As a result, unlike in traditional shape-from-shading (SfS), points with the same albedo are imaged darker the further they are, being the decrease of intensity a function of the square distance to the light source. To exploit this, we introduce a deep network, as depicted by Figure 1, that estimates depths and albedos from the image, infers normals from depths, and then renders an image while taking into account the atten-

[*] *equal contribution*

uation factor due to the distance between the light source and the surface. At training time, we minimize the difference between the original and rendered images. This enforces consistency of the depths, normals, and albedos and provides the required self-supervision without depth annotations. At inference, we use our trained network to predict depth from a RGB image and then, as our method is totally self-supervised, we can perform test-time refinement (TTR) for every monocular image, minimizing the difference between the input and rendered views, further refining the predicted depths. Our quantitative evaluation on a phantom colon dataset, where ground-truth is available, shows that our *self-supervised* approach delivers results that are very close to that of the best supervised one, and significantly superior to that of multi-view self-supervision and synthetic-to-real transfer methods. Crucially, we show quantitatively that our method keeps working on real data, for which there is no ground-truth data that can be used for training and self-supervised alternatives underperform. The main specific contributions that led to such results are 1) the inclusion of illumination decline and the endoscope's photometric calibration in the rendering equation, which provides a strong supervisory signal, and 2) a single-view self-supervised method using such renders, including two-headed network architectures LightDepth U-Net and LightDepth DPT (see details in Figure 3) that can be trained in large colonoscopy datasets without requiring ground truth labels and even further refined online in the test views.

2. Related Work

Generic Single-view Depth Estimation. It has enjoyed a renaissance after the seminal work by Eigen et al. [14], which demonstrated the effectiveness of deep neural networks for supervised pixel-wise depth regression in natural images. Subsequent research efforts have made contributions in many different directions. To name a few, network architectures evolved to fully convolutional in Laina et al. [31] and more recently to transformers [48, 4, 34]. Some of those works [4, 34] also discretize the continuous depth space into bins and formulate the problem as an ordinal regression, as in Fu et al. [18]. Other advances include interpretability [13], uncertainty quantification [46, 52], and modeling camera intrinsics [15, 21]. All these approaches are supervised and require depth ground-truth data, which can be difficult and expensive to acquire.

Self-supervised methods seek to overcome this limitation and reduce the need for ground-truth data, often by exploiting multi-view photometric consistency [19, 71, 69, 65, 29, 20]. This also enables depth refinement at test time [8, 59, 38, 58, 62, 28]. Unfortunately, this kind of supervision can be noisy, due to inaccuracies in the camera motion estimation, perspective distortions, occlusions or non-Lambertian effects, among others. As result, state-of-the-art self-supervised methods typically suffer from significantly larger inaccuracies than supervised ones. By contrast, our approach avoids these sources of errors and delivers accuracies that are close to those of supervised techniques.

Endoscopic Single-view Depth Estimation. Single-view depth estimation has been extensively studied for endoscopic purposes. Visentini et al. [61] used CT renderings for depth supervision in bronchoscopies. However, CT scans in particular and ground-truth depth data in general are very rare in endoscopy, which makes self-supervision a quasi necessity. Many works explore multi-view integration [37, 63, 25] combined with tracking and SLAM pipelines [51, 44, 39]. Others propose video-based training schemes [30, 17, 26]. Unfortunately multi-view self-supervision is even more challenging in endoscopy than in other areas due to the presence of deformations and weak texture.

Due to the specificity of the domain, synthetic to real transfer has also been extensively explored. For example, in [57] a conditional GAN is used for depth recovery while integrating SLAM and multi-view inputs. In [7], a depth network is trained with synthetic images of a simple colon model and fine-tuned with domain-randomized photorealistic images rendered from CT scans. Many other works address the domain shift between simulated and real colons [40, 41, 50, 30, 9, 53]. Learning in supervised and transferring the knowledge using uncertainty [36] uses monocular videos and multi-view stereo to provide weak depth supervision. We will show in the results section that our approach yields more accurate results, especially given that our approach to self-supervision allows further refinement of the estimates at inference time.

Shape from Shading (SfS). Depth estimation from a single image can be traced back to the early SfS methods summarized in [66] and in particular to traditional shape-from-shading [24]. However, these older techniques rely on strong assumptions that do not hold in endoscopic imagery: the camera and directional point light model are located at infinity; the reflectance is Lambertian; the albedo is constant, and the surfaces are smooth.

Importantly, lights at infinity result in ill-posed problems [47]. By contrast, when the light source is co-located with the camera that is *not* distant from the target surfaces, there is a $1/d^2$ attenuation of pixel intensity with distance d to the surface, which makes the problem well-posed when the albedo is assumed to be constant. Experimental validation that this still holds when the light source is translated with respect to the optical centre is provided in [11, 60], but still assuming constant and known albedo. Photometric stereo infers depth capturing several images from the same monocular camera under lights at different locations, but requires endoscopic hardware modifications [22, 45, 10].

More recently, the topic was revisited by SIRFS (Shape, Illumination, and Reflectance from Shading) [2] that model the interdependences between shape, illumination and reflectance, and introduces statistical priors on these quantities to disentangle their effects. In subsequent works [32, 33, 55, 35, 68], priors are learned by deep neural networks using supervision, synthetic-to-real or multi-view self-supervision. In contrast, our approach does not require such priors, which makes its deployment easier.

The SfS methods applied to endoscopy require an accurate geometrical and photometrical model of the camera and light source. This can be obtained with endoscope calibration [43, 23, 1, 3].

3. LightDepth

We use a self-supervised single-view approach to train a neural network to predict the albedo, depth, and normals at every pixel of an image so that the image can be resynthesized from these values. As shown in Fig. 1, we exploit this property using a dual-branch network that outputs pixel-wise depths and albedos. The normals are estimated analytically from the depths, and, together with the albedos, are used to render images that should be as close as possible to the original ones. At the heart of this approach is the fact that the renderer takes into account light decline as a function to distance to the surface. This is what provides the necessary self-supervisory signal.

3.1. Photometric Model

As in [3, 43], we model scene illumination as coming from a single spotlight source located at $\mathbf{x}_l \in \mathbb{R}^3$ in the camera reference frame, as depicted by Fig. 2. Spotlights usually emit with different intensities in each direction. Hence, we adopt the spotlight model (SLS) of [43]. For surface point $\mathbf{x}_i$ with off-axis angle ψ_i, we write its radiance as

$$\sigma_{\text{SLS}}(\mathbf{x}_i, \psi_i) = \frac{\sigma_0}{\|\mathbf{x}_i - \mathbf{x}_l\|^2} R(\psi_i) \ , \tag{1}$$

$$R(\psi_i) = e^{-\mu(1-\cos(\psi_i))} \tag{2}$$

where σ_0 is the maximum radiance and $R(\psi_i)$ is the radial attenuation controlled by a spread factor μ. Note that the light reaching the surface is subject to the inverse-square law and decays with the propagation distance from $\mathbf{x}_l$ to $\mathbf{x}_i$.

Light Decline. In endoscopes, the camera and the light source move jointly in a dark environment. Hence, the attenuation of the illumination is an indirect indicator of scene depth as seen from the camera. More specifically, for each pixel, we can write the rendering equation

$$\mathcal{I}(d_i, \rho_i, g) = \left(\frac{\sigma_0}{\|d_i \mathbf{r}_i - \mathbf{x}_l\|^2} R(\psi_i) \cos(\theta_i) \ \rho_i \ g \right)^{1/\gamma} , \tag{3}$$

where d_i is the depth of the i-th pixel with image coordinates $\mathbf{u}_i$, $\mathbf{r}_i = \pi^{-1}(\mathbf{u}_i)$ is the camera ray such that $\mathbf{x}_i = d_i \mathbf{r}_i$ and $\pi^{-1}(\cdot)$ is the inverse projection model of the camera. θ_i stands for the light's incidence angle with respect to the surface normal $\mathbf{n}_i$, such that, $\cos\theta_i = \mathbf{l}_i \cdot \mathbf{n}_i$. ρ_i represents the albedo of the surface at that point. g denotes the gain applied by the camera and γ is the gamma correction commonly applied by cameras to adapt images to human perception. The resulting $\mathcal{I}(d_i, \rho_i, g)$ is the color captured by the camera.

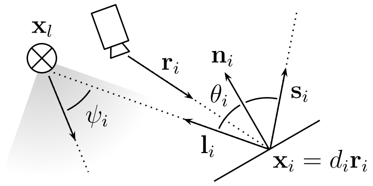

Figure 2. Spotlight illumination model, a spotlight source at position $\mathbf{x}_l$ illuminates the surface point $\mathbf{x}_i$. The emission has $R(\psi_i)$ radial fall-off, suffers from an inverse-square decline with $\mathbf{x}_l \rightarrow \mathbf{x}_i$ and attenuates with the incidence angle (θ_i). $\mathbf{l}_i$, $\mathbf{n}_i$, $\mathbf{r}_i$ and $\mathbf{s}_i$ are unit vectors.

Our model assumes Lambertian reflections, meaning that the light hitting the surface is scattered equally in all directions. The percentage of reflected light is known as albedo. Specular reflections, which are prevalent in endoscopic images, are not captured by this model but we will consider them in a specific loss that we describe in Section 3.2.

Calibration. Each endoscope has different geometric and photometric parameters, the former affecting the inverse project model π^{-1} and the latter impacting both the light position $\mathbf{x}_l$ and spread R. We can estimate these parameters for a particular endoscope by minimizing the reprojection and photometric errors on images of a calibration target, similar to [1, 3]. In our case, the auto-gain values of the endoscope are not known, so radiance measurements of the camera are unitless. Thus, we arbitrarily set $g = 1$, $\sigma_0 = 1$ and obtain up-to-scale reconstructions. Our calibration errors are between ±3 gray levels.

3.2. Self-Supervision Losses

Formally, the network of Fig. 1 takes as input an image $I \in [0,1]^{w \times h \times 3}$, estimates a depth map $\widehat{d} \in (0, \infty)^{w \times h}$ and an albedo map $\widehat{\rho} \in [0,1]^{w \times h \times 2}$. It infers normals $\widehat{\mathbf{n}}$ from $\widehat{d}$, and uses $\widehat{d}$, $\widehat{\mathbf{n}}$, and ρ to render an image $\widehat{I} \in [0,1]^{w \times h \times 3}$ that should be as similar as possible to I. To train this network, we minimize a loss

$$\mathcal{L} = \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_{sp} \mathcal{L}_{sp} \ , \tag{4}$$

where λ_s and λ_{sp} are scalar weights and $\mathcal{L}_p$, $\mathcal{L}_s$, and $\mathcal{L}_{sp}$ are the loss terms described below.

$\mathcal{L}_p$ is a photometric loss and we take it to be the squared L_2 distance between the original image I and the rendered

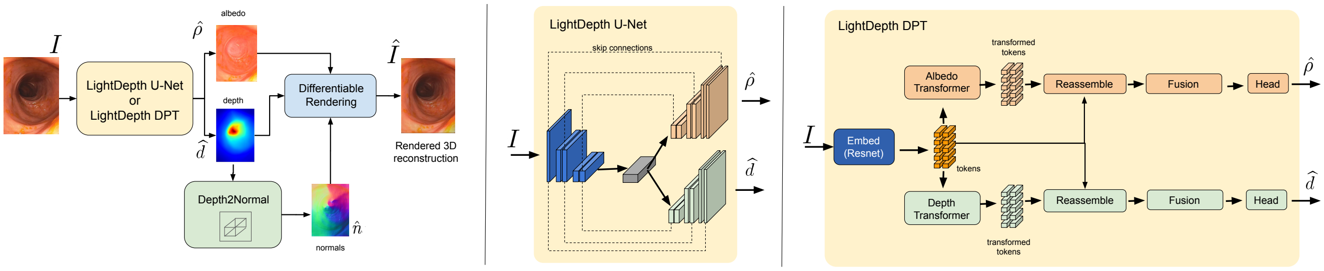

Figure 3. **Network Architecture. Left.** The input image is fed into a neural network that predicts albedo and depth values for each pixel. From the estimated depths, we compute the normals at each pixel surface using a kernel-based approach. Then, the depths, albedos, and normals are sent to a differentiable renderer that takes into account illumination decline and the endoscope's photometric model, and generates a synthetic image that should be as similar as possible to the original one. We also use specular reflections in saturated pixels to self-supervise normals. We investigated two different architectures: **Center.** LightDepth U-Net is based on a standard U-Net [54] with two decoding branches. **Right.** LightDepth DPT is based on the DPT-Hybrid architecture [48], with a second decoder branch added for the albedo.

one $\hat{I}$. Note that because our rendering model is fully differentiable, we can perform end-to-end training.

$$\mathcal{L}_p = \sum_{i \in \Omega} (I_i - \widehat{I}_i)^2, \quad \text{where} \quad \widehat{I}_i = \mathcal{I}(i, \widehat{d}_i, \widehat{\rho}_i, g) \tag{5}$$

As in [20], $\mathcal{L}_s$ is a regularization term that minimizes depth gradients except in areas of high color gradients, that may correspond to depth discontinuities. We write

$$\mathcal{L}_s = |\partial_x \widehat{d}| e^{-|\partial_x I|} + |\partial_y \widehat{d}| e^{-|\partial_y I|} \tag{6}$$

Finally, recall that we made a Lambertian assumption in Eq. 3, which prevents us to account properly for specular reflections and the overexposed pixels they produce. This is a potential source of error and fails to exploit the very useful information that specularities provide about normals. To remedy this, we introduce specular loss $\mathcal{L}_{sp}$. Given image location i, the corresponding direction $\mathbf{l}_i$ from the surface to the light source and the normal of a the surface $\widehat{\mathbf{n}}$, the law of reflection states that

$$\mathbf{s}_i = \mathbf{l}_i - 2\hat{\mathbf{n}}_i \left(\hat{\mathbf{n}}_i \cdot \mathbf{l}_i \right) \tag{7}$$

is the specularly reflected direction. Hence, we take our specular loss term to be

$$\mathcal{L}_{sp} = \sum_{i \in \Omega} \left(m_i \left(\mathbf{s}_i \cdot (-\mathbf{r}_i) - 1 \right) \right)^2 \ , \tag{8}$$

$$m_i = \begin{cases} 1 & I_i > th \\ 0 & \text{otherwise} \end{cases}$$

which minimizes the discrepancy between the expected specular reflection $\mathbf{s}_i$ and the actual direction $(-\mathbf{r}_i)$ where the camera observes the reflection, resulting in pixel with high intensity $th = 0.98$.

Our method takes a single image as input, which makes 3D shape recovery solely from pixel colors an underconstrained problem. According to Eq. 3, a change in the brightness of a pixel can be due to changes in depth, albedo, camera exposure or surface normal. For example, if a given pixel is very bright, it can be because the pixel is close to the camera/light; the surface has a different albedo, resulting in more light being reflected; the surface normal is aligned to the light/camera, which increases the reflected light; the camera exposure and digital gain have been increased, which impacts brightness values in the whole image. Given the albedo at each surface point and the camera auto-gain, we could resolve these ambiguities. However, in medical endoscopy, true albedos are unknown, and auto-gain is not provided by the hardware manufacturer.

Albedo Constancy. We observe that endoscopy images exhibit a limited range of colors, with brighter tones being present in close areas and darker tones in deeper regions. Consequently, we hypothesize a significant correlation between albedo and the chromatic attributes, namely Hue and Saturation, in the HSV color space, as well as between depth and the Value Channel. In this way, we constrain the palette of colors that can be explained by the albedo decoder and we enhance the disentanglement between depth and albedo by setting $V = 100$ for all albedo values. Hence, to predict the albedo map $\widehat{\rho}$, our network predicts just two channels per pixel, for Hue and Saturation, and assumes Value to be one to convert to the RGB space, in which the loss is formulated.

3.3. Network Architecture

Our network outputs depth and albedo maps. In Fig. 3, we provide a more detailed depiction of our encoder-decoder architecture. We have tested two different versions. The first one is a U-net with two decoders and skip connections, with a ResNet18 serving as the backbone. Our decoders design is inspired by [20]. The second one relies on visual transformers for depth estimation [48]. As a backbone, we use a Resnet-50 (DPT-Hybrid) and two decoders that reassemble the tokens and apply attention heads. Further details regarding these architectures can be found in the supplementary material.

In both versions, to compute the normals at any given pixel, we use a convolution kernel with six-neighborhood (N, NE, E, S, SW, and W) in the depth map. We define six triangles using the central pixel as reference, with each triangle having its own normal. The normal of the central pixel is computed as the average of the normals of the triangles weighted by their area. The use of six neighbors lets us reuse triangles during the convolution pass to speed up computation.

4. Results

4.1. Datasets

We evaluated LightDepth and relevant baselines on three endoscopy datasets: An in-house *synthetic colon*, *C3VD* [5], and *EndoMapper* [1]. With these, we can show quantitative and qualitative results with several levels of realism. **Synthetic Colon.** We simulate a real Olympus CF-H190L endoscope consisting on a fish-eye camera and a spot-light source, both calibrated as in [1]. This is in contrast to other synthetic datasets that simulate arbitrary camera and illumination configurations, typically pinhole cameras with no or arbitrary distortion and ideal light sources with no radial falloff. [50, 67, 44, 49], We rendered the images using ray-casting techniques, in which the colon's geometry and albedo are defined by a triangle mesh obtained from a CT scan of a real colon [27]. We ignore global illumination effects and assume Lambertianity, so there are no specular reflections. The influence of these two effects will be assessed in the two other datasets. Our synthetic data is hence composed by 1620 fish-eye RGB frames annotated with per-pixel albedo, depth and normals. We split it into 1168 images for training and 452 images for test. Example frames can be found in the supplementary material.

C3VD [5] contains real images recorded in a phantom with ground-truth depth. The images have been captured by a real Olympus CF-HQ190L endoscope in a phantom silicone model of a human colon. The data is annotated with ground-truth depth and normals by applying 2D-3D registration of the 3D phantom models. The authors claim that the silicone material is opaque, hence we can assume that the only light source available is in the endoscope. Finally, it includes a geometrical calibration based on the Scaramuzza model [56]. C3VD provides a good compromise between realism (real endoscope, global illumination effects and specular highlights) and ground-truth labels for quantitative evaluation. Of the 10,088 images available, we use 7,200 for training and 2,888 for testing. In the supplementary material, we provide the sections of the phantom used for testing and training.

EndoMapper [1] provides the most challenging data, as it contains real colonoscopy and gastroscopy procedures inside the human body, performed by endoscopists on a day-to-day basis. Here we find real textures such as veins, blood and dirt, and other effects such as blur, water and frames very close or even hitting the mucosa. Foam and bubbles are indeed very common in endoscopy images and are usually ignored. LightDepth is capable of disentangling these as part of the albedo and not of the depth. Before processing the dataset, we perform a manual inspection of the selected sequences and we eliminate occluded and excessively blurred frames.

Finally, we train in three procedures, consisting of two colonoscopies and one gastroscopy. There are a total of 24,444, 23,456 and 3,032 frames, respectively. Details of the sequences and frames we use can be found in the supplementary material.

4.2. Metrics, Baselines, and Training Details

We report results using a median-based scale alignment for all methods, even those supervised with real-scale depth, for fairness. In our experiments, we compare against models that use depth supervision and multi-view self-supervision. For depth supervision, we use two different architectures, U-Net with L1 loss as a representative of convolutional architectures and DPT-Hybrid [48] as a state-of-the-art representative of transformer-based models, learning inverse depth with an scale invariant loss.

For a fair comparison, we also evaluate our LightDepth using the same U-Net and DPT architectures. The U-Net is pre-trained on ImageNet dataset [12]. For DPT, we initialize with the author-provided weights for encoder and depth decoder. The albedo decoder is trained from scratch. During training, we select a smoothing weight $\lambda_s = 0.1$ in Eq. 4 and a learning rate of 10^{-4} for the Adam optimizer. In the synthetic dataset, we trained our network with λ_{sp} = 0, as synthetic dataset has no specular reflections. In C3VD and EndoMapper, we use $\lambda_{sp} = 1$.

Test-Time Refinement (TTR). As our LightDepth enables single-view self-supervision, we can continuously refine the depth predictions online, obtaining much more accurate reconstructions. In the results denoted as "(TTR)", we perform online test-time refinement for each test image separately during $N = 20$ optimization steps, using the loss $\mathcal{L}$ in Equation 4, as in training time. To mitigate the risk of catastrophic forgetting, we load again the original model trained in the train split after TTR for each image.

Note in Table 1 how TTR improves significantly the metrics with respect to LightDepth without TTR for U-Net and

							Depth [mm]						Normals [°]
Dataset	Architecture	Backbone	Supervision	MAE ↓	MedAE ↓	RMSE ↓	RMSE_{log} ↓	Abs_{Rel} ↓	Sq_{Rel} ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	MAE ↓
Synthetic	U-Net	ResNet18	Depth GT	**4.37**	2.99	**6.38**	**0.1251**	0.0965	**0.0008**	0.9057	**0.9931**	**0.9997**	25.1
	LightDepth U-Net	ResNet18	Light	4.76	**2.47**	8.60	0.1375	**0.0903**	0.0011	**0.9180**	0.9820	0.9935	**15.2**
C3VD	U-Net	ResNet18	Depth GT	4.15	3.29	5.52	0.1139	0.0902	0.0007	0.9172	0.9943	**0.9994**	26.5
	DPT-Hybrid [48]	ResNet50	Depth GT	**3.22**	2.77	**4.10**	**0.0860**	**0.0699**	**0.0004**	**0.9640**	0.9865	0.9913	**15.1**
	Monodepth2 [20]	ResNet50	Multi-View	14.27	9.59	18.64	0.3921	0.2971	0.0070	0.4897	0.7313	0.8611	43.6
	CADepth [64]	ResNet18	Multi-View	52.35	17.04	87.43	0.9144	1.1916	0.2650	0.3664	0.5653	0.6679	67.2
	XDCycleGAN [42]	ResNet	Cycle	17.16	11.91	22.43	0.4953	0.3616	0.0105	0.4291	0.6615	0.7910	64.4
	LightDepth U-Net	ResNet18	Light	4.37	2.92	6.31	0.1183	0.0856	0.0007	0.9315	0.9934	**0.9994**	24.0
	LightDepth DPT	ResNet50	Light	3.94	2.67	5.60	0.1080	0.08046	0.0006	0.9476	0.9965	**0.9994**	21.3
	LightDepth U-Net	ResNet18	Light (TTR)	3.72	2.59	5.43	0.1060	0.0770	0.0005	0.9505	**0.9971**	**0.9994**	23.5
	LightDepth DPT	ResNet50	Light (TTR)	3.70	**2.58**	5.27	0.1073	0.0780	0.0005	0.9525	0.9961	0.9992	22.5

Table 1. Depth and normal metrics for several architectures and supervision modes. Best results per dataset are bolfaced, second best underlined.

Dataset	Architecture	Supervision	SSIM ↑	MAE ↓
Synthetic	LightDepth U-Net	Light	0.9901	0.0192
C3VD	LightDepth U-Net	Light	0.9765	0.0657
	LightDepth DPT	Light	0.8873	0.0599
	LightDepth U-Net	Light (TTR)	**0.9811**	**0.0276**
	LightDepth DPT	Light (TTR)	0.8977	0.0329

Table 2. SSIM and MAE for rendered images in C3VD. Test-time refinement (TTR) gives a substantial improvement.

DPT architectures. Remarkably, observe how TTR even outperforms the metrics achieved by Depth GT supervision. Figure 4 shows the improvement given by TTR in the network prediction of depth, normals and albedo and overall in the 3D reconstruction. Inference time is $\sim$ 5ms for LightDepth U-Net and $\sim$ 22 ms for LightDepth DPT on a NVIDIA GeForce RTX 3090. We can do TTR in $\sim$ 90 ms per optimization step in U-Net and $\sim$ 190 ms in DPT.

4.3. Quantitative Results on Synthetic and Phantom

Synthetic colon. The first two rows in Table 1 report depth and normal metrics for a U-Net supervised with Depth GT, and our self-supervised LightDepth U-Net architecture. Observe that the metrics are similar. This is notable, as self-supervision is consistently reported in the literature to underperform with respect to depth supervision, and suggests that illumination decline provides a very strong self-supervisory signal in endoscopies, which our experiments in the other two datasets confirm.

Furthermore, light self-supervision outperforms Depth GT supervision in MedAE and $\delta < 1.25$, which means that most of the error distribution is lower for light self-supervision and only a small fraction of large errors are better with depth supervision. We observed that it is in far and dark areas where light self-supervision is weaker and this produces a higher depth MAE and RMSE. Observe the significantly lower error in normal with our light self-supervision, due to the lower errors in most pixels.

C3VD Phantom. We report depth and normal metrics on the real phantom images of the C3VD dataset in Table 1. Our self-supervised architectures LightDepth U-Net and LightDepth DPT with TTR outperform supervision with

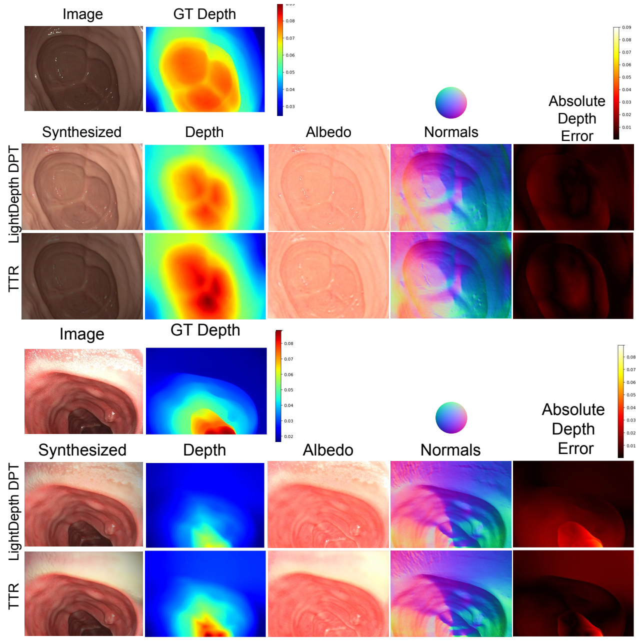

Figure 4. DepthLight and DepthLight TTR on C3VD. Our light decline captures the correct shape of the cecum in the first image and the shape of the polyp in the second. Note how the estimates of normals and albedo are similar before and after TTR. By optimising depth by reducing illumination, DepthLight achieves a darker appearance and improvements in depth estimation.

Depth GT in MedAE, while the rest of the metrics are very close. As in the case of the synthetic dataset, this is a remarkable result because self-supervised architectures typically lag behind supervised ones in single view depth estimation. The fact that LightDepth MedAE is better and RMSE is worse suggests that our errors are better in most of the distribution, and there are a few regions with large errors where Depth GT supervision is able to offer an advantage. Table 2 details metrics on the quality of the rendered image, which suggest the strength of the self-supervision signal. Observe the improvement of this metrics for the TTR case.

In Table 1, observe that the multi-view self-supervised baselines, Monodepth2 [20] and CADepth [64], have a poor performance in our data, worse in comparison than results in other datasets. This could be due to the weak textures and changing lighting in the colonoscopy images, resulting in noisy estimations for relative motion and uninformative photometric residuals. Being single-image, our approach is impervious to such difficulties.

Domain shift. As synthetic-to-real is common in endoscopies to address the lack of ground-truth depth for supervision, we also evaluated XDCycleGAN [42] as a baseline. Note that the domain shift is still affecting the results. Our single-view LightDepth self-supervision enables training in the target domain, and hence removes completely the domain shift, achangedchieving significantly lower errors.

Dataset			Depth [mm]			Normals [°]
Train	Test	Supervision	MAE ↓	MedAE ↓	RMSE↓	MAE ↓
Synt.	Synt.	Depth GT	4.37	2.99	6.38	25.1
		Light	4.76	2.47	8.60	15.2
Synt.	C3VD	Depth GT	9.44	5.79	12.83	73.7
		Light	5.09	3.51	7.14	27.7
		Depth GT (TTR)	4.96	3.14	7.11	25.4
		Light (TTR)	3.80	**2.51**	5.54	23.6
C3VD	C3VD	Depth GT	4.15	3.29	5.52	26.5
		Light	4.37	2.92	6.31	24.0
		Light (TTR)	**3.72**	2.59	**5.43**	**23.5**

Table 3. Synthetic-to-real domain shift. Best results in C3VD test set are boldfaced, second best are underlined. Note the domain shift effect between Synt. and C3VD test data in the bigger errors, and how TTR removes the domain shift effect completely. Notably, our LightDepth TTR delivers similar errors than the models without domain shift, trained in C3VD.

Table 3 elaborates further on domain shift by showing depth and normals metrics for a U-Net architecture in these cases. Specifically, we trained a U-Net model with Depth GT supervision and light self-supervision in our synthetic dataset and evaluated their performance in the synthetic and C3VD test sets. Observe how the domain shift affects all metrics significantly. Interestingly, the model trained with light self-supervision and without TTR generalizes significantly better to the C3VD data, as our LightDepth self-supervised model is closer to the physical phenomena than Depth GT supervision. Again, note that single-view self-supervision removes completely the domain shift effect, as models can be trained directly in the target domain. Very remarkably, the performance of our models with domain shift after TTR matches the performance of the models without domain shift.

Normals from Depth. The literature details different manners to obtain surface normals from a depth map, e.g., [16, 6]. Table 4 shows a MAE analysis of the most promising ones in C3VD. Specifically, we evaluate four methods: a U-net trained to regress normals from depth, the recent TFtN method [16], the implementation in Open3D [70] that computes normals from a k-nearest neighbourhood in the point cloud, and an in-house method that uses six-neighbourhood in the image. Our analysis shows that an analytic average in a neighbourhood is significantly better than a U-Net and TFtN, and our in-house method that considers a neighbourhood in the image is slightly better, so this last one was our choice.

Method	MAE [°]
U-Net	16.24
TFtN [16]	3.89
Open3D [70]	1.67
In-house	**1.32**

Table 4. Normal's MAE for baseline methods.

Ablation Study on the Loss. In Table 5, we ablate the terms of our loss function. The smoothness prior ($\mathcal{L}_s$ term) is remarkably beneficial for both depth and normal prediction. When we do not take advantage of the information of the specular reflections (no $\mathcal{L}_{sp}$ term), we obtain worse results. Adding this new loss term, we see how all the depth and normal metrics improve, especially in the median error, which outperforms the supervised and now matches that obtained in the simulation experiment. Still, the depth MAE and RMSE are slightly higher than those of the baseline due to the far spurious points.

	Depth [mm]			Color	Normals [°]
Loss	MAE ↓	MedAE ↓	RMSE↓	MAE ↓	MAE ↓
$\mathcal{L}_p$	6.05	3.93	8.79	0.0637	35.5
$\mathcal{L}_p + \mathcal{L}_s$	4.95	3.04	7.23	0.0690	24.6
$\mathcal{L}_p + \mathcal{L}_s + \mathcal{L}_{sp}$	4.37	2.92	6.31	0.0657	24.0

Table 5. Ablation study of the losses with LightDepth U-Net in C3VD dataset. Observe the improvement given by each term.

4.4. Qualitative Results in Real Endoscopy

We now turn to real images of a human colon from the EndoMapper dataset and present qualitative results in Figure 5. Additional ones can be found in the supplementary material. Some details are recovered very accurately, such as the normal maps showing clearly the tubular shape; the depth maps reflecting the discontinuities in the Haustras; the albedos capturing the blood vessels, in particular in the 5th column; and the bubbles and fluids colors in the 6th and 7th columns, which make the 3d reconstruction of these bubbles and fluids very plausible.

Unfortunately, there is no ground-truth data available for this dataset, which prevent us from presenting quantitative results, and we do not know of any other dataset with real colonoscopy images that includes ground-truth data. Nevertheless, visual inspection of our results hints that the strengths of our techniques demonstrated quantitatively in Section 4.3 will carry over on truly realistic scenarios like this one.

5. Limitations and Discussion

As mentioned in Section 3.2, our depth predictions are up-to-scale. Even if the camera auto-gain was available, the albedo scale may be challenging to learn, so estimating the real scale is not straightforward. In any case, other methods such as multi-view self-supervision or synthetic-to-real cannot guarantee an accurate estimation of the scale either. We assume that Lambertian reflectance is prevalent in most

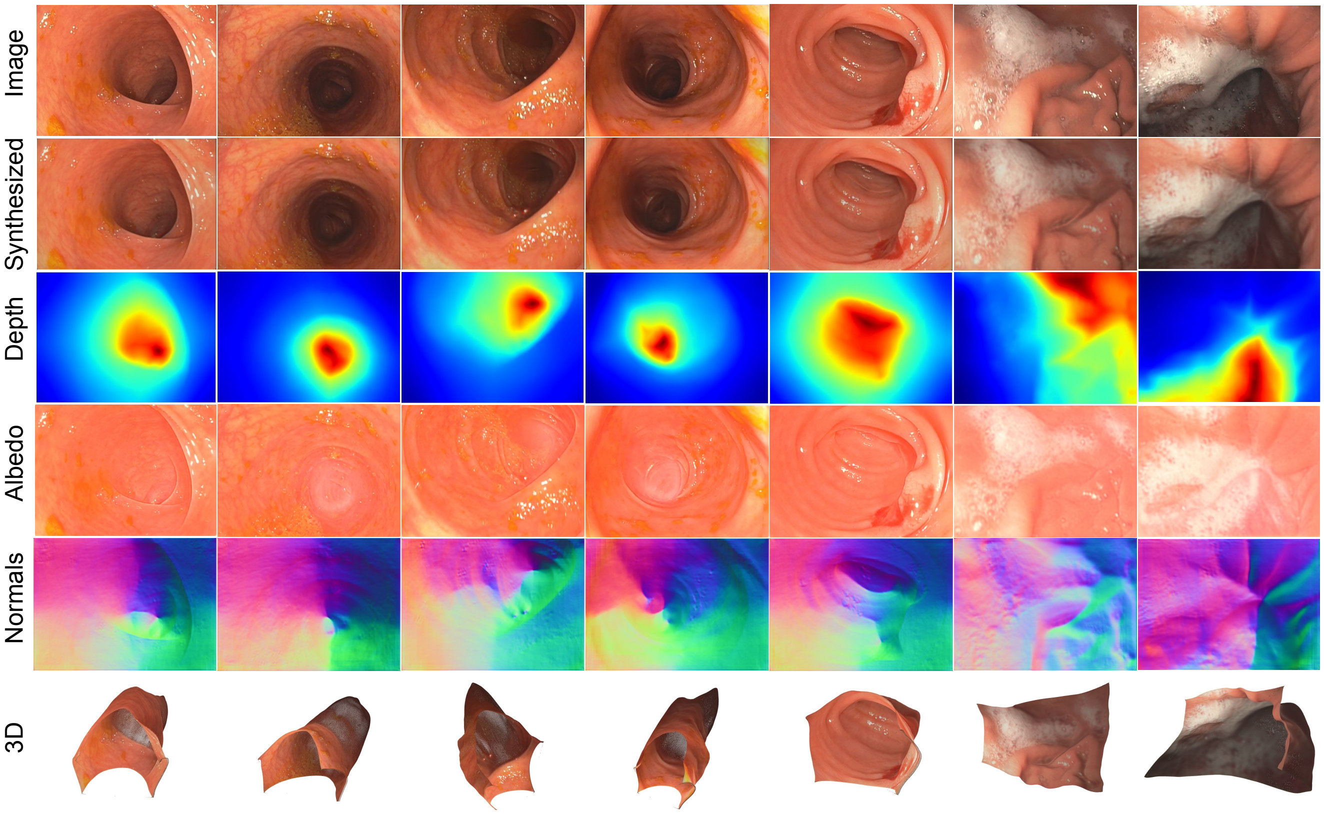

Figure 5. Qualitative results on EndoMapper with LightDepth DPT. Columns 1–5 are real colonoscopy images, and columns 6–7 are real gastroscopy images. In colonoscopies, observe that the normals exhibit a tubular shape specific of the colon. The albedo prediction captures disruptions such as veins, blood, dirt, foam and specularites. Note the influence of light decline in the image and the correlation with the estimated depths.

tissues, and for areas where this does not hold, we use a basic model to capture specularities. Further research could focus on the application of more sophisticated photometric models that cover specularities, e.g., the Phong model.

Thanks to our priors on albedo and depth, we successfully disentangle both factors in our experiments. However, our $V = 100$ prior might not hold in areas of clotted blood or with very dark albedos, e.g., because of a disease. These priors might need to be tuned in new application domains for enhanced performance. Finally, although we demonstrate this technology in the context of endoscopy, its principles are applicable in any setup in which the only light source is close to the target surface and rigidly attached to the camera. In other words, our LightDepth has the potential to open research avenues in many other domains.

6. Conclusions

In this work, we have proposed, for the first time, a single-view self-supervision method for depth learning, which we denote LightDepth, that exploits and is limited to the case of a single spotlight source co-located with a monocular camera, a case that includes, among others, the relevant application of medical endoscopy. As our main contribution, we developed the specific self-supervised learning setup that models the quadratic light decline and enables self-supervised learning. We have implemented two different architectures, a first one based on convolutions and a second one based on transformers, and evaluated their performance against ground-truth supervision, multi-view self-supervision, and domain transfer approaches. Our results show that LightDepth outperforms multi-view self-supervision and synthetic-to-real transfer and matches the performance of fully supervised approaches. Not only that, its training and test-time refinement setup is significantly simpler: LightDepth only requires a reasonable endoscope calibration and does *not* require camera motion estimation nor ground-truth labels nor realistic simulations, all of them challenging in endoscopies. This unlocks, from a practical point of view, relevant potential applications in the medical domain.

Acknowledgements

This work was supported by the EU Comission (EU-H2020 EndoMapper GA863146), the Spanish

Government (PID2021-127685NB-I00, FPU20/0678, PID2021-125209OB-I00, PGC2018-096367-B-I00 and TED2021-131150BI00) and the Aragon Government (DGA-T45_23R).

References

[1] Pablo Azagra et al. EndoMapper dataset of complete calibrated endoscopy procedures. *arXiv:2204.14240*, 2022. 3, 5

[2] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. 3

[3] Víctor M. Batlle, José M.M. Montiel, and Juan D. Tardós. Photometric single-view dense 3D reconstruction in endoscopy. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4904–4910, 2022. 3

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4009–4018, 2021. 2

[5] Taylor L. Bobrow, Mayank Golhar, Rohan Vijayan, Venkata Akshintala, Juan R. Garcia, and Nicholas J. Durr. Colonoscopy 3D video dataset with paired depth from 2D-3D registration. *arXiv:2206.08903*, 2022. 5

[6] Alexandre Boulch and Renaud Marlet. Deep learning for robust normal estimation in unstructured point clouds. In *Computer Graphics Forum*, volume 35, pages 281–290. Wiley Online Library, 2016. 7

[7] Richard J Chen, Taylor L Bobrow, Thomas Athey, Faisal Mahmood, and Nicholas J Durr. SLAM endoscopy enhanced by adversarial depth prediction. *KDD Workshop on Applied Data Science for Healthcare*, 2019. 2

[8] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *IEEE/CVF International Conference on Computer Vision*, pages 7063–7072, 2019. 2

[9] Kai Cheng, Yiting Ma, Bin Sun, Yang Li, and Xuejin Chen. Depth estimation for colonoscopy images with self-supervised learning from videos. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, 2021. 2

[10] Toby Collins and Adrien Bartoli. 3D reconstruction in laparoscopy with close-range photometric stereo. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 634–642. Springer, 2012. 2

[11] Toby Collins and Adrien Bartoli. Towards live monocular 3D laparoscopy using shading and specularity information. In *Int. Conf. Inf. Process. in Computer-Assisted Interventions*, pages 11–21. Springer, 2012. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[13] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2366–2374, Cambridge, MA, USA, 2014. MIT Press. 2

[15] José M. Fácil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11826–11835, 2019. 2

[16] Rui Fan, Hengli Wang, Bohuan Xue, Huaiyang Huang, Yuan Wang, Ming Liu, and Ioannis Pitas. Three-filters-to-normal: An accurate and ultrafast surface normal estimator. *IEEE Robotics and Automation Letters*, 6(3):5405–5412, 2021. 7

[17] Daniel Freedman, Yochai Blau, Liran Katzir, Amit Aides, Ilan Shimshoni, Danny Veikherman, Tomer Golany, Ariel Gordon, Greg Corrado, Yossi Matias, et al. Detecting deficient coverage in colonoscopies. *IEEE Transactions on Medical Imaging*, 39(11):3451–3462, 2020. 2

[18] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018. 2

[19] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[20] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *IEEE/CVF International Conference on Computer Vision*, October 2019. 2, 4, 5, 6

[21] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019. 2

[22] Yang Hao, Jing Li, Fei Meng, Peisen Zhang, Gastone Ciuti, Paolo Dario, and Qiang Huang. Photometric stereo-based depth map reconstruction for monocular capsule endoscopy. *Sensors*, 20(18):5403, 2020. 2

[23] Yang Hao, Marco Visentini-Scarzanella, Jing Li, Peisen Zhang, Gastone Ciuti, Paolo Dario, and Qiang Huang. Light source position calibration method for photometric stereo in capsule endoscopy. *Advanced Robotics*, 34(12):789–801, 2020. 3

[24] Berthold K.P. Horn and Michael J. Brooks, editors. *Shape from Shading*. MIT Press, 1989. 2

[25] Baoru Huang, Jian-Qing Zheng, Anh Nguyen, David Tuch, Kunal Vyas, Stamatia Giannarou, and Daniel S Elson. Self-supervised generative adversarial network for depth estimation in laparoscopic images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, 2021. 2

[26] Seung-Jun Hwang, Sung-Jun Park, Gyu-Min Kim, and Joong-Hwan Baek. Unsupervised monocular depth estimation for colonoscope system using feedback network. *Sensors*, 21(8), 2021. 2

[27] Kağan İncetan, Ibrahim Omer Celik, Abdulhamid Obeid, Guliz Irem Gokceler, Kutsev Bengisu Ozyoruk, Yasin Almalioglu, Richard J Chen, Faisal Mahmood, Hunter Gilbert, Nicholas J Durr, et al. VR-Caps]: a virtual environment for capsule endoscopy. *Medical image analysis*, 70:101990, 2021. 5

[28] Sergio Izquierdo and Javier Civera. Sfm-ttr: Using structure from motion for test-time refinement of single-view depth networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[29] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4756–4765, 2020. 2

[30] Mert Asim Karaoglu, Nikolas Brasch, Marijn Stollenga, Wolfgang Wein, Nassir Navab, Federico Tombari, and Alexander Ladikos. Adversarial domain feature adaptation for bronchoscopic depth estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 300–310. Springer, 2021. 2

[31] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Fourth IEEE International Conference on 3D Vision (3DV)*, pages 239–248, 2016. 2

[32] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. In *Computer Graphics Forum*, volume 37, pages 409–419, 2018. 3

[33] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2475–2484, 2020. 3

[34] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. BinsFormer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 2

[35] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W Jacobs. Shape and material capture at home. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6123–6133, 2021. 3

[36] Xingtong Liu, Ayushi Sinha, Masaru Ishii, Gregory D. Hager, Austin Reiter, Russell H. Taylor, and Mathias Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Transactions on Medical Imaging*, 39(5):1438–1447, 2020. 2

[37] Huoling Luo, Qingmao Hu, and Fucang Jia. Details preserved unsupervised depth estimation by fusing traditional stereo knowledge from laparoscopic images. *Healthcare Technology Letters*, 6(6):154, 2019. 2

[38] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020. 2

[39] 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. 2

[40] F. Mahmood, Richard Chen, and Nicholas J Durr. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Transactions on Medical Imaging*, 37(12):2572–2581, 2018. 2

[41] Faisal Mahmood and Nicholas J Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Medical image analysis*, 48:230–243, 2018. 2

[42] Shawn Mathew, Saad Nadeem, Sruti Kumari, and Arie Kaufman. Augmenting colonoscopy using extended and directional CycleGAN for lossy image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4696–4705, June 2020. 6, 7

[43] Richard Modrzejewski, Toby Collins, Alexandre Hostettler, Jacques Marescaux, and Adrien Bartoli. Light modelling and calibration in laparoscopy. *Int. J. Computer Assisted Radiology and Surgery*, 15(5):859–866, 2020. 3

[44] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, et al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis*, 71:102058, 2021. 2, 5

[45] Vicente Parot, Daryl Lim, German Gonzalez, Giovanni Traverso, Norman S. Nishioka, Benjamin J. Vakoc, and Nicholas J. Durr. Photometric stereo endoscopy. *Journal of Biomedical Optics*, 18(7):076017, 2013. 2

[46] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3227–3237, 2020. 2

[47] Emmanuel Prados and Olivier Faugeras. Shape from shading: a well-posed problem? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 870–877, 2005. 2

[48] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021. 2, 4, 5, 6

[49] Anita Rau, Binod Bhattarai, Lourdes Agapito, and Danail Stoyanov. Bimodal camera pose prediction for endoscopy. *arXiv preprint arXiv:2204.04968*, 2022. 5

[50] Anita Rau, PJ Eddie Edwards, Omer F Ahmad, Paul Riordan, Mirek Janatka, Laurence B Lovat, and Danail Stoyanov. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–10, 2019. 2, 5

[51] David Recasens, José Lamarca, José M Fácil, JMM Montiel, and Javier Civera. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4), 2021. 2

[52] Javier Rodríguez-Puigvert, Rubén Martínez-Cantín, and Javier Civera. Bayesian deep neural networks for supervised learning of single-view depth. *IEEE Robotics and Automation Letters*, 7(2):2565–2572, 2022. 2

[53] Javier Rodriguez-Puigvert, David Recasens, Javier Civera, and Ruben Martinez-Cantin. On the uncertain single-view depths in colonoscopies. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, pages 130–140, Cham, 2022. Springer Nature Switzerland. 2

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[55] Shen Sang and Manmohan Chandraker. Single-shot neural relighting and svbrdf estimation. In *European Conference Computer Vision (ECCV)*, pages 85–101. Springer, 2020. 3

[56] Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5695–5701, 2006. 5

[57] Mali Shen, Yun Gu, Ning Liu, and Guang-Zhong Yang. Context-aware depth and pose estimation for bronchoscopic navigation. *IEEE Robotics and Automation Letters*, 4(2), 2019. 2

[58] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference Computer Vision (ECCV)*, pages 572–588. Springer, 2020. 2

[59] Lokender Tiwari, Pan Ji, Quoc-Huy Tran, Bingbing Zhuang, Saket Anand, and Manmohan Chandraker. Pseudo RGB-D for self-improving monocular SLAM and depth prediction. In *European Conference Computer Vision (ECCV)*, pages 437–455. Springer, 2020. 2

[60] Marco Visentini-Scarzanella, Danail Stoyanov, and Guang-Zhong Yang. Metric depth recovery from monocular images using shape-from-shading and specularities. In *IEEE International Conference on Image Processing*, pages 25–28, 2012. 2

[61] Marco Visentini-Scarzanella, Takamasa Sugiura, Toshimitsu Kaneko, and Shinichiro Koto. Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. *International Journal of Computer Assisted Radiology and Surgery*, 12(7):1089–1099, Jul 2017. 2

[62] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, 2021. 2

[63] Ke Xu, Zhiyong Chen, and Fucang Jia. Unsupervised binocular depth prediction network for laparoscopic surgery. *Computer Assisted Surgery*, 24(sup1):30–35, 2019. 2

[64] Jiaxing Yan, Hong Zhao, Penghui Bu, and YuSheng Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *IEEE International Conference on 3D vision (3DV)*, pages 464–473, 2021. 6

[65] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, 2018. 2

[66] Ruo Zhang, Ping-Sing Tsai, J.E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999. 2

[67] Shuai Zhang, Liang Zhao, Shoudong Huang, Menglong Ye, and Qi Hao. A template-based 3D reconstruction of colon structures and textures from stereo colonoscopic images. *IEEE Transactions on Medical Robotics and Bionics*, 3(1):85–95, 2020. 5

[68] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18643–18652, 2022. 3

[69] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. DeepTAM: Deep tracking and mapping. In *European Conference Computer Vision (ECCV)*, pages 851–868. Springer, 2018. 2

[70] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 7

[71] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017. 2