Photometric single-view dense 3D reconstruction in endoscopy

Víctor M. Batlle, J.M.M. Montiel, *Member, IEEE* and Juan D. Tardós, *Fellow, IEEE*

***Abstract*— Visual SLAM inside the human body will open the way to computer-assisted navigation in endoscopy. However, due to space limitations, medical endoscopes only provide monocular images, leading to systems lacking true scale. In this paper, we exploit the controlled lighting in colonoscopy to achieve the first in-vivo 3D reconstruction of the human colon using photometric stereo on a calibrated monocular endoscope. Our method works in a real medical environment, providing both a suitable in-place calibration procedure and a depth estimation technique adapted to the colon's tubular geometry. We validate our method on simulated colonoscopies, obtaining a mean error of 7% on depth estimation, which is below 3 mm on average. Our qualitative results on the EndoMapper dataset show that the method is able to correctly estimate the colon shape in real human colonoscopies, paving the ground for true-scale monocular SLAM in endoscopy.**

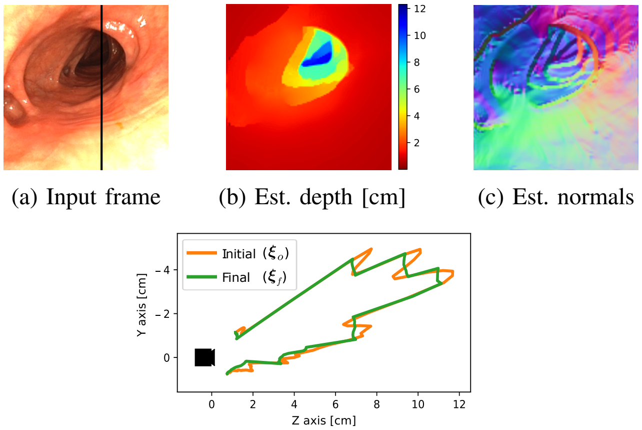
I. INTRODUCTION

With the goal of improving the efficiency and effectiveness of routine diagnostic and medical intervention procedures, there is a growing research interest in extending augmented reality and autonomous navigation to the human body. These advances will need to accurately solve localization and mapping from visual sensors.

Simultaneous Localization and Mapping (SLAM) with stereo [1] and visual-inertial [2] cameras already provide great accuracy for multi-view reconstruction. In most medical endoscopy applications, space limitations restrict to monocular cameras. With monocular vision, the real scale of the environment cannot be observed, so potential applications are limited to up-to-scale reconstructions which usually suffer from scale drift problems, specially in deforming environments [3].

However, the interior of the human body is an example of an artificially illuminated environment, where the light source is controlled and linked to the camera movement. Our goal is to take advantage of the illumination to reconstruct 3D scenes, obtaining photometric stereo information by considering a light-camera pair.

The main contributions of this paper are: (1) a simple photometric model for the endoscope light and camera, (2) a photometric calibration method that does not require a Lambertian pattern and can be carried out in-place on a hospital setting, and (3) the first method capable of reconstructing the geometry of the human colon from a single view, only from the illumination of a calibrated conventional monocular endoscope (see Fig. 1). This would solve the scale-drift problem in monocular SLAM, and can also provide real-scale maps when the albedo of the surface and the endoscope's auto-gain are known.



(a) Input frame (b) Est. depth [cm] (c) Est. normals

(d) Cross-section along black line

Fig. 1: Depth estimation on an *in-vivo* human colonoscopy.

II. RELATED WORK

Recent results in single-view depth estimation using deep convolutional networks [4] open the possibility of designing accurate SLAM systems from monocular cameras, specially hybrid approaches [5] which combine deep learning with traditional methods. Previous work demonstrates that, using a depth estimation network, it is possible to perform scale-aware monocular SLAM, obtaining almost the same accuracy as with stereo, and eliminating scale drift [6], [7], [8], [9].

A first attempt to apply these methods to endoscopy sequences achieves real-scale reconstructions with good accuracy [10]. However, these methods require stereo supervision to learn how to predict the true size. Thus, today its application in colonoscopy is limited by the impossibility of acquiring stereo images of the human colon.

Other authors focus on the study of photometry to obtain dense and semi-dense reconstructions of outdoor and indoor 3D scenes [11], [12]. They assume constant illumination, usually ambient light, ignoring any change in lighting.

In contrast, inside the human body, the illumination is controlled and light moves together with the camera. Recent work [13], [14] shows that changes in lighting, instead of being ignored, can be used to our benefit, obtaining dense reconstructions from monocular sequences.

* The authors are with the Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, María de Luna 1, 50018 Zaragoza, Spain. E-mail: {vmbatlle, josemari, tardos}@unizar.es.

This work was supported by EU-H2020 grant 863146: ENDOMAPPER, Spanish government grant PGC2018-096367-B-I00 and FPU scholarship of V. M. Batlle, and by Aragón government grant DGA_T45-17R.

A. Lighting model

Previous works propose a lighting model for their working environment. Specifically, they model light emission, interaction with surfaces, and capture by the camera. So far, the complexity of these *ad-hoc* models required them to be calibrated and tested in laboratory environments. In this paper, we propose a simplified model that allows easier calibration without the need for Lambertian patterns.

Modrzejewski et al. [13] do a thorough work on analyzing various light source models. Their Spot Light Source (SLS) model offers a good compromise between complexity and accuracy. We adopt a similar approach, but with the aim of modeling the multiple light sources of the endoscope as a single virtual light. This leads us to a generic model in which our virtual light is located at the camera's optical center.

Hao et al. [15] calibrate the light emission separately by means of a plane mirror. Conversely, we propose a joint calibration method, which also allows easy estimation of camera geometry and photometry at the same time.

A common approach [13], [16] consists of assuming Lambertian surfaces, both during calibration and during reconstruction inside the human body. However, we show that this causes bias in the calibration when the real surface is not perfectly Lambertian. If not corrected, this error propagates to the 3D reconstruction. In contrast, our calibration considers non-Lambertian properties, giving results that are not affected by the calibration pattern used.

B. Reconstruction

In photometric stereo, the discussed light model can be used to obtain a dense depth map of the scene. The main discrepancy between the different approaches lies in their corresponding reconstruction method.

Modrzejewski et al. [13] propose an initial multi-view reconstruction followed by a photometric optimization, where a regularization term tends to favor smooth planar surfaces. Crucially, the multi-view method requires a rigid environment. In contrast, our method is based only on lighting, being able to reconstruct the environment from a single view. In addition, the geometry of the human colon, with numerous discontinuities, is far from planar. By considering this in our regularization, we can reconstruct its complex shape.

Hao et al. [14] focus on specular highlights, where their method achieves the best accuracy. However, the accuracy of their reconstruction decreases for the rest of the continuous surface. Unlike them, we perform a global optimization, considering each point equally, which does not require the surface to be continuous.

To the best of our knowledge, previous work on dense photometric reconstruction on endoscopy [13], [14], [17], [18] has been validated on nearly planar scenes, without discontinuities. Focusing on colonoscopy, Parot et al. [19] provide experimental validation on phantoms. Instead, we demonstrate that our method can recover for the first time the tubular topology of a human colon, form a single *in-vivo* video frame, preserving the anatomical folds of the intestine, known as haustra.

III. ENDOSCOPE MODEL

This paper presents a photometric approach to the problem of monocular 3D reconstruction during medical endoscopy. This approach considers the geometric model of image formation and the photometric model of light transport (Fig. 2).

A. Geometric model

An endoscope camera is designed to cover a wide view angle. Thus, we adopt the approach of Kannala & Brandt [20] to model the fisheye lens, with four projective parameters f_x, f_y, C_x, C_y, and four distortion coefficients k_{1-4}. We denote the optical center of the camera as $\mathbf{x}_\mathrm{c}$.

B. General photometric model

The illumination system mounted on an endoscope usually consists of one or more small lights. Given its small size, these lights are usually modeled as punctual lights [13], [14]. Thus, radiance L_i coming from light center $\mathbf{x}_l$ to a point $\mathbf{x}$ in a surface, is subject to the inverse-square law:

$$L_\mathrm{i}(\mathbf{x}) = \mu(\mathbf{x}) \frac{\sigma_o}{\|\mathbf{x} - \mathbf{x}_l\|^2} \tag{1}$$

In most endoscopes light is transmitted to the tip using optical fiber. Therefore, the amount of light emitted in each direction of space is not uniform. We model this behavior with a light spread function $\mu(\mathbf{x})$, that can be specified by fixing a principal direction $\mathbf{d}_l$, over which maximum radiance σ_o is emitted, and assuming a radial cosine fall-off [14]. We decide to modulate this decay by adding the cosine exponent k as a parameter:

$$\mu(\mathbf{x}) = \cos^k \psi, \quad \psi = \langle \mathbf{x} - \mathbf{x}_l,\ \mathbf{d}_l \rangle \tag{2}$$

When light reaches a surface, most of it will be reflected, going out in different directions depending on the material properties. The Bidirectional Reflectance Distribution Function (BRDF) $f_\mathrm{r}(\omega_\mathrm{i}, \omega_\mathrm{r})$ defines how light is reflected at an opaque surface. Usually, this behavior depends on the incoming ω_i and outgoing ω_r direction of the light ray with respect to the normal $\mathbf{n}$ of the surface at that point. The inclination of the incident ray modifies the area of the projection of the solid angle on the surface, depending on the cosine of its angle with the normal. As a result, the reflected radiance is:

$$L_\mathrm{r}(\mathbf{x}, \omega_\mathrm{r}) = L_\mathrm{i}(\mathbf{x}, \omega_\mathrm{i})\, f_\mathrm{r}(\omega_\mathrm{i}, \omega_\mathrm{r}) \cos\theta \tag{3}$$

where $\theta = \langle \omega_\mathrm{i}, \mathbf{n} \rangle$ and, for our case, ω_i is the direction to the light source and ω_r points to the camera (see Fig. 2).

Light reaching the camera is affected by a set of factors. The capture system usually introduces attenuation on the received radiance. Natural vignetting tends to approximate to $\cos^4 \alpha$, where α is the off-axis angle between the ray direction and the camera forward $\mathbf{z}$ [21]. Mechanical vignetting is not easy to model theoretically. Therefore, vignetting is usually empirically approximated [22]. We assume radial attenuation from the camera's forward vector, by modeling the decay with a k' exponent on a cosine function:

$$V(\mathbf{x}) = \cos^{k'} \alpha, \quad \alpha = \langle \mathbf{x} - \mathbf{x}_\mathrm{c},\ \mathbf{z} \rangle \tag{4}$$

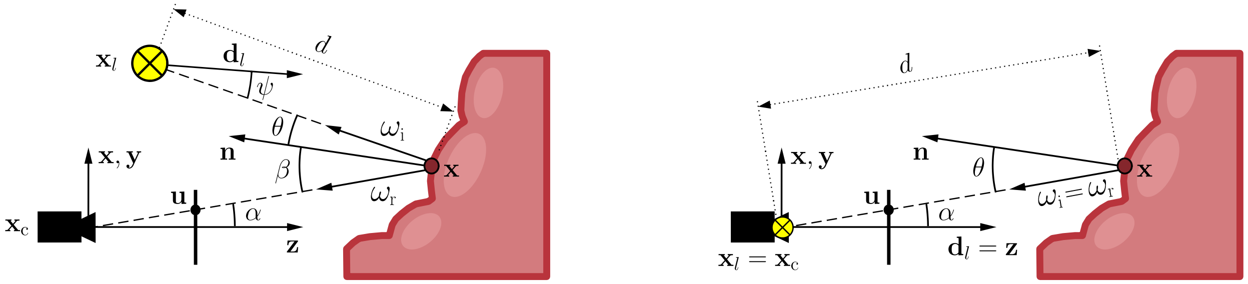


Fig. 2: **Left:** General photometric model. **Right:** Simplified photometric model. We assume a virtual light is located at the camera optical center and the light's principal direction is at the camera forward vector.

Endoscope cameras might automatically adjust some parameters, such as exposure time or signal amplification. In video streams, this is controlled by an automatic gain control (AGC) logic. We assume this auto-gain usually acts as a multiplying factor g_t at each t-th time instant. In order to increase the perceived dynamic range, cameras map the captured values through a gamma function with a common value of $\gamma = 2.2$ that does not change over time.

Our complete photometric model considers all concepts introduced above, as a combination of light, surface and camera effects:

$$\mathcal{I}(\mathbf{x}) = \left(\frac{\mu(\mathbf{x})\, \sigma_\mathrm{o}}{\|\mathbf{x} - \mathbf{x}_l\|^2}\; f_\mathrm{r}(\omega_\mathrm{i}, \omega_\mathrm{r})\; \cos\theta\; V(\mathbf{x})\; g_\mathrm{t} \right)^{1/\gamma} \tag{5}$$

C. Simplified photometric model

The presented model takes into account only one light source. Each additional light source must be modeled independently, in a similar way, but adding complexity to the model. Moreover, the characteristics of each endoscope version vary slightly. Commonly we find two or three optical fiber guides, which conduct the light to different points on the tip of the endoscope.

Instead of the costly process of modeling the details of each specific hardware, we propose a simplification based on encapsulating the joint effects of all the light points into a single virtual light source (see Fig. 2). We observe that these light points are usually distributed fairly symmetrically around the endoscope camera. Therefore, we decided to place the virtual light source at the optical center of the camera, i.e. $\mathbf{x}_l = \mathbf{x}_\mathrm{c}$ and we align light's principal direction with the camera forward, i.e. $\mathbf{d}_l = \mathbf{z}$.

In this new set-up, camera's vignetting and virtual light's spread function are coupled, i.e. $\psi = \alpha$. Thus, we model the effect of both functions jointly, as:

$$\mu'(\mathbf{x}) = \mu(\mathbf{x})\, V(\mathbf{x}) = \cos^k \alpha \tag{6}$$

Regarding surface reflectance, now incoming ω_i and reflected ω_r directions match. Thus, the domain of the BRDF can be simplified. We will consider the incident angle θ of the light on the surface, such that the BRDF is simply $f_\mathrm{r}(\theta)$.

In most endoscopes auto-gain logic is unknown. Therefore, g_t values are coupled with the absolute radiance σ_o, so that their effects cannot be separated. Consequently, we fix the σ_o parameter to an arbitrary value and estimate the relative auto-gain changes.

Finally, we obtain a simplified photometric model, which is parameterized according to the unknowns we want to estimate for our endoscope:

$$\mathcal{I}\left(\mathbf{x},\, k,\, g_\mathrm{t},\, \gamma\right) = \left(\frac{\mu'(\mathbf{x},\, k)}{\|\mathbf{x} - \mathbf{x}_l\|^2}\; f_\mathrm{r}(\theta)\;\; \cos\theta\;\, g_\mathrm{t} \right)^{1/\gamma} \tag{7}$$

IV. ENDOSCOPE CALIBRATION

Geometric and photometric calibration is performed with a small Vicalib [23] pattern printed on a white paper sheet of 5.61×9.82 cm. From a video captured with the endoscope, geometric parameters are obtained by processing 1 out of 20 frames using the Vicalib software.

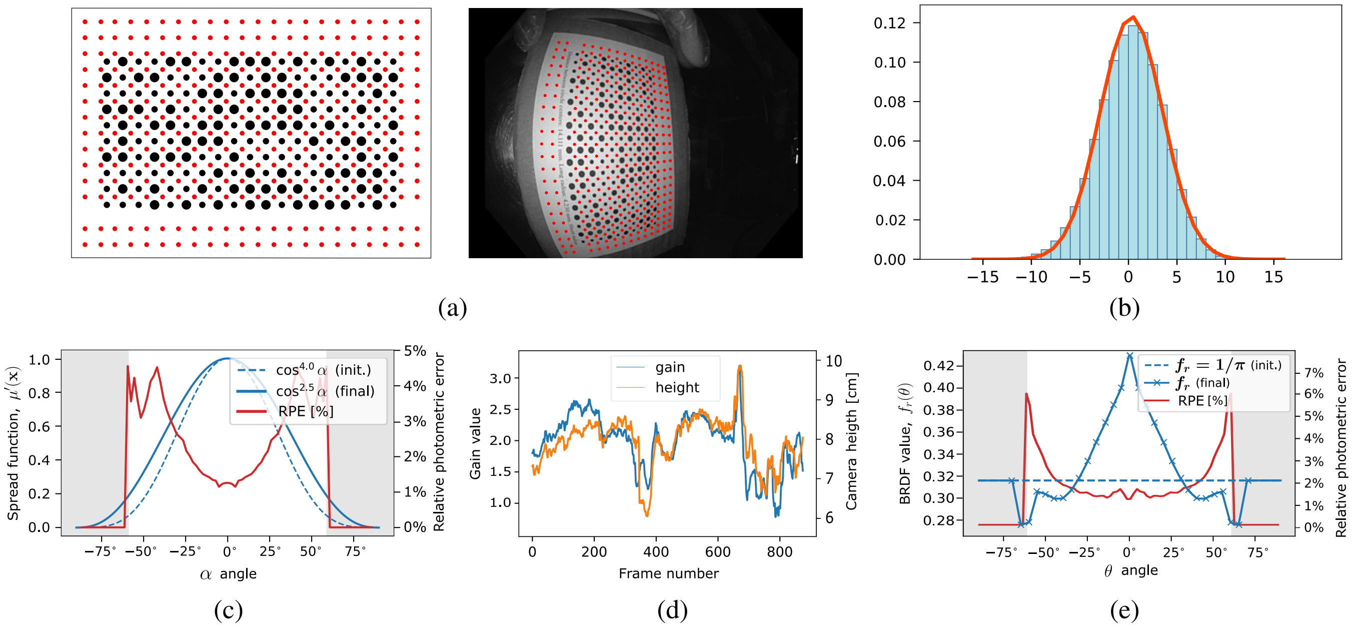
Focusing on photometry, we propose an optimization problem that aims to minimize the photometric error between the empirical data I and our model. For that, we select a set of j sample points, uniformly distributed along the white areas of the pattern (see Fig. 3a). Then, the photometric loss function is computed on every visible $\boldsymbol{x}_\mathrm{j}$ point on each t-th frame of the video, using Huber function ρ to be robust against spurious:

$$\{k, g_\mathrm{t}, \gamma \mid \forall \mathrm{t}\}^* = \underset{k,\, g_\mathrm{t},\, \gamma}{\operatorname{argmin}} \sum_{\mathrm{j,t}} \rho\left(I_\mathrm{jt} - \mathcal{I}\left(\boldsymbol{x}_\mathrm{j}, k, g_\mathrm{t}, \gamma\right)\right) \tag{8}$$

A. Results

From a video resolution of 1440×1080 px at 30 frames per second, the calibration obtains values for the geometric parameters $f_x = 717.21$ px, $f_y = 717.48$ px, $C_x = 735.37$ px, $C_y = 552.80$ px and the four distortion coefficients k_{1-4} are $[-0.13893,\, -1.2396E{-}03,\, 9.1258E{-}04,\, -4.0716E{-}05]$. These lead to a reprojection error in RMSE of 0.5288 px.

Regarding the photometric calibration results, the optimization converges to $k = 2.5$, $\gamma = 2.2$ and estimates auto-gain g_t values ranging from 1 to 3. We observe that the final spread is wider than the natural vignetting $\cos^4 \alpha$ (see Fig. 3c). This is consistent with the illumination system of our endoscope, where three real light sources result in a widening of our virtual light's spread. Validation results show a small relative error of 1% in the center, i.e. 0°. However,



(a)

(b)

(c)

(d)

(e)

Fig. 3: Sampling the Vicalib pattern: (a) Red marks correspond to each $\boldsymbol{x}_j$ sampled point. Photometric calibration results: (b) Photometric errors of the calibrated model are close to a Gaussian distribution with a mean of 0.3 and std. of 3.2 gray levels. (c) Joint attenuation caused by light spread function $\mu(\mathbf{x})$ and camera vignetting $V(\mathbf{x})$. (d) Estimated auto-gain factors over the calibration video. (e) Non-Lambertian BRDF for the paper sheet used for calibration.

error grows towards the edges, reaching 4.5% at 40° and when $\alpha > 60°$ (shaded area) the function remains unsampled in our calibration data.

Automatic gain cannot be evaluated with test data, because each image has a different gain factor. Instead, we can see that the estimated gain factor follows a continuous progression over time along the calibration sequence (see Fig. 3d). Moreover, the gain value for each frame seems to be closely related to the distance from the camera to the illuminated surface. That is, when the camera is closer to the pattern, light is more intense, and the endoscope applies a lower gain value.

In addition, we observed that modeling the paper reflectance with a Lambertian BRDF $f_r(\theta) = 1/\pi$ led to a biased calibration. Therefore, we decided to also optimize the value of the BRDF for fifteen values of the θ angle and apply linear interpolation in the rest of the domain of the function (see Fig. 3d). The estimated BRDF for the paper sheet shows specular behavior when the camera is close to the perpendicular to the surface. This results in a peak in the reflectance when the θ angle is near zero.

The result of the calibration allows us to estimate the gray level of a pixel with a standard deviation of 3.2 levels. The distribution of errors is unbiased (see Fig. 3b). Moreover, the new estimated BRDF is an isolated component of the model. Therefore, when we want to apply our calibration in the interior of the human body, we can replace this BRDF with that of the human colon, and the rest of the calibrated parameters remain valid.

V. Depth Estimation

Given the calibrated endoscope photometric model and a single endoscope image, our goal is to estimate depth and surface normal for each imaged 3D point. We consider the following assumptions:

- Similarly to Modrzejewski et al. [13], we assume that human tissue can be approximated by a Lambertian material if specular highlights are masked or treated as spurious. For this, we propose an automatic method for highlight detection and inpainting.
- In addition, given the weak texture of the colon tissue, the surface albedo k_d is measurable and considerably constant. In our experiments we set $f_r(\theta) = k_d/\pi$.
- The imaged surfaces are smooth, except at occasional discontinuities. This allows us to approximate differential changes of the surface by a tangent plane.

Based on DTAM method [11], we approach the estimation of a depth map as an optimization problem, that minimizes an energy function:

$$E_{\boldsymbol{\xi}} = \int_{\Omega} \Big\{ C(\mathbf{u}, \boldsymbol{\xi}(\mathbf{u})) + \lambda R(\mathbf{u}, \boldsymbol{\xi}) \Big\} d\mathbf{u} \qquad (9)$$

where

- $\mathbf{u} \in \Omega$ are coordinates on the image,
- $\boldsymbol{\xi} : \Omega \rightarrow \mathbb{R}$ is the depth map,
- $C(\,)$ is a photometric cost function,
- $R(\,)$ is a regularization cost,
- $\lambda \in \mathbb{R}^+$ adjusts the regularization weight.

A. Photometric cost function

DTAM assumes ambient light on the scene and uses a cost function based solely on camera geometry and brightness constancy. However, the illumination during endoscopy varies with camera movement. Consequently, we replace the original cost function with a novel cost function based on our photometric endoscope model:

$$C(\mathbf{u}, d) = \rho\left(I(\mathbf{u}) - \mathcal{I}\left(\pi^{-1}(\mathbf{u}, d)\right)\right) \tag{10}$$

where

- d is the Euclidean distance to the world point,
- $\pi^{-1}(\,)$ is the camera unprojection model,
- $\mathcal{I}(\,)$ is our calibrated endoscope photometric model (7),
- $I : \Omega \to \mathbb{R}^+$ denotes the actual pixel intensity,
- $\rho(\,)$ is the Huber robust cost function.

B. Normal estimation from a depth map

The photometric model of a scene $\mathcal{I}$ is influenced by both the distance to the points (inverse-square law) and the surface normal (cosine term). However, surface normal is directly related to depth variations. Therefore, both parameters should not be optimized separately. Instead, given the local planarity assumption, we can calculate the normal of a point from the estimated depth map [24].

Thanks to this relationship, we keep the depth map as the only unknown variable of the problem. However, it should be noted that this method is influenced by spurious data, especially at surface discontinuities. Therefore, in these areas, we expect some localized errors.

C. Smoothness regularization

The defined cost function is trying to find three unknowns per pixel (d, n_θ, n_φ) from one intensity measurement (I). In order to solve the problem's ill-posedness, DTAM proposes a regularization term that penalizes local depth variations, except at points where the luminosity gradient is large, which usually correspond to surface discontinuities:

$$R(\mathbf{u}, \boldsymbol{\xi}) = g(\mathbf{u}) \|\nabla \boldsymbol{\xi}(\mathbf{u})\|_\epsilon \tag{11}$$

Thus, $\|x\|_\epsilon$ Huber norm with $\epsilon \approx 1.0e^{-4}$ works as total variation (TV) regularizer and $g(\mathbf{u})$ reduces the regularization strength at high gradient points. Thanks to these two terms, (d, n_θ, n_φ) are now constrained by the pixel's neighborhood, and at the same time the discontinuities of the colon can be preserved.

However, this might not be the best regularizer in the colon's tubular geometry. The first derivative $\nabla\boldsymbol{\xi}$ always favors zero changes along the depth map. So the reconstruction will tend to a plane parallel to the camera. Instead, similarly to [13], we use the second-order derivative $\nabla^2\boldsymbol{\xi}$ to impose smoothness, although we continue to allow discontinuities.

D. Depth map representation

DTAM formulates its depth map $\boldsymbol{\xi}_{1/z}$ as the inverse distance in the Z-axis. This decision is appropriate for a multi-view-based problem, as the pinhole projection model depends directly on this variable. Instead, we are faced with a single-view problem. In our case, the photometry is quadratically dependent on the inverse of the Euclidean distance, i.e. $\mathcal{I} \propto 1/d^2$.

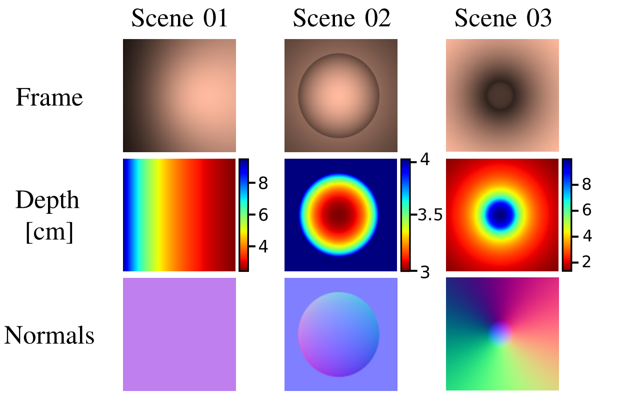


Fig. 4: Data generated for our simple geometry dataset. **Top:** Frame simulated with our model as in (7). **Middle:** Ground-truth Z-depth map. **Bottom:** Ground-truth normal map, represented in color space $(R, G, B) = ([n_x, n_y, n_z] + 1)/2$.

Therefore, we will compare the previous formulation with two new variants of the depth map, such that

$$\boldsymbol{\xi}_{1/z} = \frac{1}{z}, \qquad \boldsymbol{\xi}_d = d, \qquad \boldsymbol{\xi}_{1/d} = \frac{1}{d} \tag{12}$$

E. Initial solution

We make the optimization method start from an initial solution, where we assume all surface normal vectors pointing towards the camera optical center.

From the calibrated photometric model, we revert the effects of light spread function, Lambertian BRDF, as well as known camera gain and gamma correction:

$$I_\mathrm{c}(\mathbf{u}) = \frac{I(\mathbf{u})^\gamma}{\mu'(\mathbf{x}) \cdot f_\mathrm{r}(\theta) \cdot g_\mathrm{t}} = \frac{\cos\theta}{d^2} \tag{13}$$

where I_c is a *canonical intensity value*, which is obtained after compensating all mentioned parameters that influence image formation. Note that, when a surface normal points towards the camera, the θ angle is zero. Therefore, by solving for d in the above equation, we get an initial solution

$$d_o(\mathbf{u}) = I_\mathrm{c}(\mathbf{u})^{-1/2} \tag{14}$$

The closer the actual θ is to zero, the closer this initial solution is to the real depth.

VI. EXPERIMENTAL RESULTS

In this section, we first conduct some simple experiments to determine the best smoothness regularizer and depth map variant. Then, we check the accuracy of our depth estimation method with a photo-realistic simulation of a human colon. Finally, we test our method in a real in-vivo colonoscopy image.

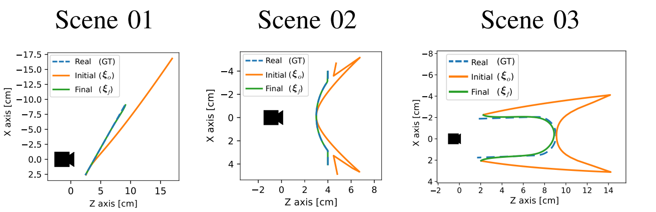


Fig. 5: Cross-section along Y-axis (middle of the image) of actual surface (GT), initial solution (ξ_o) and final optimized depth map (ξ_f). Our estimation matches the ground-truth.

TABLE I: ACCURACY ON SIMPLE GEOMETRY DATASET

Scene	ξ	Reg.	# iter.	Depth error [mm]		Depth error [%]		Normals error [deg]	
				Mean	Median	Mean	Median	Mean	Median
01	$1/z$	∇	1 148	1.0	**<0.1**	0.90	**0.04**	1.19	0.20
		∇^2	**73**	**0.3**	**<0.1**	**0.32**	0.09	**0.62**	**0.18**
02	$1/z$	∇	102	0.2	0.2	0.35	0.37	1.00	0.50
		∇^2	**64**	**0.1**	**0.1**	**0.25**	**0.21**	**0.95**	**0.39**
03	$1/z$	∇	>5 550	3.0	2.1	7.30	6.73	12.17	9.44
		∇^2	>1 500	**1.9**	**1.8**	5.83	5.31	**11.07**	**8.00**
	d	∇^2	301	**1.9**	**1.8**	5.85	**4.99**	12.92	9.15
	$1/d$	∇^2	**78**	**1.9**	**1.8**	**5.78**	5.21	11.55	8.30

A. *Simple geometry dataset*

The first experiment consists of a simple simulation, based on our photometric model as in (7). We simulate a rotated plane, a curved surface, and a tubular geometry (see Fig. 4).

Table I presents the results of this experiment. We conclude that the regularizer of the second derivative is better for accuracy (see in Fig. 5 the best result for each scene) and is also faster in convergence.

Regarding the depth map variants, in a tubular geometry, inverse Euclidean distance $\xi_{1/d}$ performs better than the alternatives. The corresponding results in Table I show much faster convergence with similar accuracy.

B. *Simulated colon dataset*

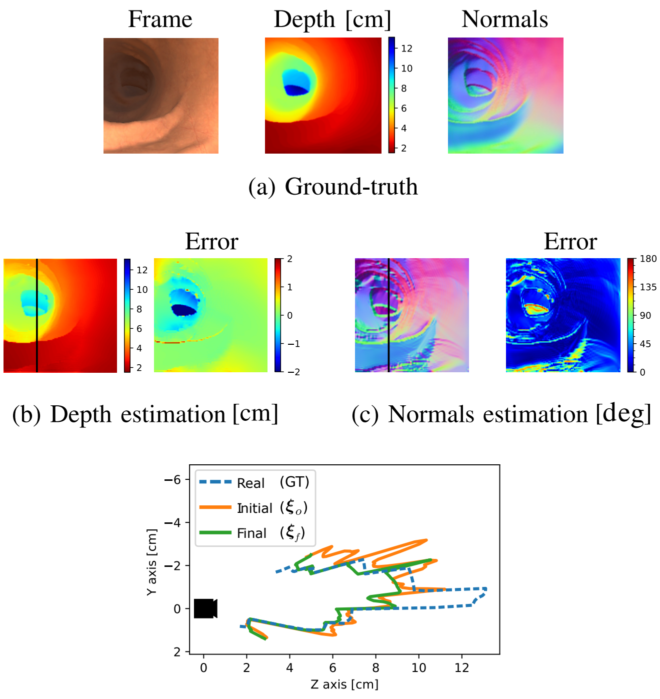
In this experiment, we validate our method on a frame of a photo-realistic dataset [25]. This dataset simulates an endoscopy procedure based on a real CT scan of a human colon. The simulation includes effects more similar to those found in a real environment, such as richer textures and ambient light caused by secondary reflections, that are not considered in our model.

This input frame comes from an endoscope without distortion or vignetting and in which the light spread is homogeneous (see Fig. 6a). We also know the average albedo of the surface and the gain of the endoscope. The results of this experiment are shown in the Table II.

As in the previous case, we see that $\xi_{1/d}$ is the best variant of the method, providing a good estimation (see Fig. 6b) with less than 3 mm error on average. However, on this photo-realistic simulation, the ∇^2 regularizer obtains lower accuracy. The second derivative is less robust to noise, such as that introduced by surface texture, which causes the albedo to be not perfectly uniform.

TABLE II: ACCURACY ON SIMULATED COLON [25]

ξ	Reg.	# iter.	Depth error [mm]		Depth error [%]		Normals error [deg]	
			Mean	Median	Mean	Median	Mean	Median
$1/z$	∇	44 500	5.3	2.3	10.60	7.41	30.63	23.77
	∇^2	8 900	5.1	4.0	15.09	11.86	36.06	29.26
d	∇	20 000	3.3	**1.6**	7.90	**4.98**	**26.21**	**18.75**
	∇^2	44 500	3.8	2.0	9.57	6.37	32.00	23.56
$1/d$	∇	44 500	**2.8**	**1.6**	**7.32**	5.01	27.89	19.69
	∇^2	**5 400**	5.1	4.0	15.16	12.05	35.86	28.89



(a) Ground-truth

(b) Depth estimation [cm]

(c) Normals estimation [deg]

(d) Cross-section along black line

Fig. 6: Results on the simulated colon dataset [25]. Using variants $\xi_{1/d}$ and $\nabla\xi$. They show good accuracy for the estimated Z-depth map and the corresponding normals.

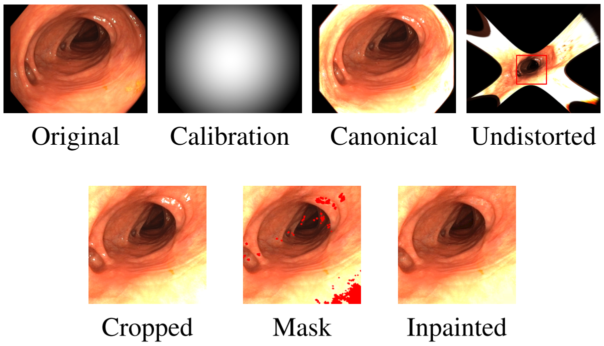


Fig. 7: HCULB real colon dataset from the EndoMapper project. We apply our photometric and geometric calibration and we perform highlight inpainting.

In addition, the photo-realistic simulator introduced a fog effect in areas far away from the camera. This increases the intensity of distant pixels. As a result, we cannot reconstruct the deepest part of the colon, from 10 to 12 cm (see Fig. 6d). Therefore, the median error of 1.6 mm is considerably lower than the mean, which is influenced by those spurious.

C. Real colon dataset

Our method is finally validated on a real image from the HCULB colonoscopy dataset (EndoMapper EU-H2020 project). The images of this dataset correspond to real human colons and are acquired *in-vivo* during medical procedures, with the endoscope we calibrated in Section IV.

We take a single image (see Fig. 7). First, we compensate for calibrated vignetting and light spread function and obtain a frame with canonical illumination. Then, we undistort and crop the image. Finally, we perform automatic highlight detection and inpainting. With this, we obtain a frame similar to the one of the previous simulated dataset.

Fig. 1 shows the reconstruction provided by our method. The estimated scale is arbitrary, as we do not have data about the automatic gain. Moreover, the HCULB dataset does not provide ground-truth information for comparison. Nevertheless, qualitatively, the result we have obtained properly reconstructs the tubular topology of the colon and also recovers notably the shape of the haustra.

VII. Conclusions

This paper proposes a photometric stereo method that is able to reconstruct for the first time the geometry of the human colon using only the illumination on real monocular endoscopy procedures. We can recover the true scale of the environment if the surface albedo and the endoscope's autogain are known. The latter is set by the manufacturer of the hardware and the albedo could be reasonably estimated in the future since it is mostly uniform along the colon.

Our method obtains reconstructions with a mean accuracy below 3 mm on simulated data and is able to reconstruct the tubular geometry on a real colon, where it preserves the discontinuities at the colon's haustra.

In addition, a calibration process is designed to suit a medical endoscope. Our experiments show that we are able to model a real endoscope with an error of 3 gray levels. This allows us to conclude that our model, based on a virtual light source, offers a good compromise between accuracy and ease of calibration in a real environment.

Currently, depth estimation works in an off-line mode, but it allows us to overcome the lack of 3D perception inherent in monocular camera systems. In this way, for example, our method could constitute a new source of self-supervision for learning depth estimation without the need for stereo.

In conclusion, these results leave the door open to future work in real-time SLAM and autonomous navigation inside the colon, solving scale drift and allowing true scale maps.

Acknowledgment

We acknowledge the work of Pablo Azagra (Universidad de Zaragoza) in performing the calibration sequence. We also thank Anita Rau (University College London) for providing preliminary data from her simulated dataset [25].

References

[1] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[2] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Trans. Robot.*, 2021.

[3] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel, "DefSLAM: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Trans. Robot.*, vol. 37, no. 1, pp. 291–303, 2020.

[4] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3828–3838.

[5] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1278–1289.

[6] Y. Li, C. Xie, H. Lu, X. Chen, J. Xiao, and H. Zhang, "Scale-aware monocular slam based on convolutional neural network," in *IEEE Int. Conf. on Information and Automation (ICIA)*, 2018, pp. 51–56.

[7] L. Tiwari *et al.*, "Pseudo rgb-d for self-improving monocular slam and depth prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 437–455.

[8] V. M. Batlle and J. D. Tardós Solano, "Scale estimation in monocular orb-slam2 using deep convolutional networks," BA thesis. University of Zaragoza, 2020.

[9] C. Campos and J. D. Tardós, "Scale-aware direct monocular odometry," *arXiv preprint*, 2021.

[10] D. Recasens, J. Lamarca, J. M. Fácil, J. Montiel, and J. Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7225–7232, 2021.

[11] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2320–2327.

[12] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, 2018.

[13] R. Modrzejewski, T. Collins, A. Hostettler, J. Marescaux, and A. Bartoli, "Light modelling and calibration in laparoscopy," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 5, pp. 859–866, 2020.

[14] Y. Hao, J. Li, F. Meng, P. Zhang, G. Ciuti, P. Dario, and Q. Huang, "Photometric stereo-based depth map reconstruction for monocular capsule endoscopy," *Sensors*, vol. 20, no. 18, p. 5403, 2020.

[15] Y. Hao, M. Visentini-Scarzanella, J. Li, P. Zhang, G. Ciuti, P. Dario, and Q. Huang, "Light source position calibration method for photometric stereo in capsule endoscopy," *Advanced Robotics*, vol. 34, no. 12, pp. 789–801, 2020.

[16] M. Visentini-Scarzanella and H. Kawasaki, "Simultaneous camera, light position and radiant intensity distribution calibration," in *Image and Video Technology*. Springer, 2015, pp. 557–571.

[17] T. Okatani and K. Deguchi, "Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center," *Computer vision and image understanding*, vol. 66, no. 2, pp. 119–131, 1997.

[18] T. Collins and A. Bartoli, "Towards live monocular 3d laparoscopy using shading and specularity information," in *Int. Conf. Inf. Process. in Computer-Assisted Interventions*. Springer, 2012, pp. 11–21.

[19] V. Parot *et al.*, "Photometric stereo endoscopy," *Journal of biomedical optics*, vol. 18, no. 7, p. 076017, 2013.

[20] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1335–1340, 2006.

[21] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[22] J. Engel, V. Usenko, and D. Cremers, "A photometrically calibrated benchmark for monocular visual odometry," *arXiv preprint*, 2016.

[23] A. R. P. G., "Vicalib visual-inertial calibration suite," 2016.

[24] S. Hinterstoisser *et al.*, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 858–865.

[25] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *International journal of computer assisted radiology and surgery*, vol. 14, no. 7, pp. 1167–1176, 2019.