frontiers | Frontiers in Robotics and AI

TYPE Technology and Code
PUBLISHED 25 July 2023
DOI 10.3389/frobt.2023.1212525

Check for updates

OPEN ACCESS

EDITED BY
Emanuele Lindo Secco,
Liverpool Hope University,
United Kingdom

REVIEWED BY
Hang Su,
Fondazione Politecnico di Milano, Italy
Chao Hu,
Zhejiang University, China

*CORRESPONDENCE
Luigi Manfredi,
✉ mail@luigimanfredi.com

RECEIVED 26 April 2023
ACCEPTED 26 June 2023
PUBLISHED 25 July 2023

CITATION
Mathew A, Magerand L, Trucco E and Manfredi L (2023), Self-supervised monocular depth estimation for high field of view colonoscopy cameras.
Front. Robot. AI 10:1212525.
doi: 10.3389/frobt.2023.1212525

Self-supervised monocular depth estimation for high field of view colonoscopy cameras

Alwyn Mathew[1], Ludovic Magerand[2], Emanuele Trucco[2] and Luigi Manfredi[1]*

[1]Division of Imaging Science and Technology, School of Medicine, University of Dundee, Dundee, United Kingdom, [2]Discipline of Computing, School of Science and Engineering, University of Dundee, Dundee, United Kingdom

Optical colonoscopy is the gold standard procedure to detect colorectal cancer, the fourth most common cancer in the United Kingdom. Up to 22%–28% of polyps can be missed during the procedure that is associated with interval cancer. A vision-based autonomous soft endorobot for colonoscopy can drastically improve the accuracy of the procedure by inspecting the colon more systematically with reduced discomfort. A three-dimensional understanding of the environment is essential for robot navigation and can also improve the adenoma detection rate. Monocular depth estimation with deep learning methods has progressed substantially, but collecting ground-truth depth maps remains a challenge as no 3D camera can be fitted to a standard colonoscope. This work addresses this issue by using a self-supervised monocular depth estimation model that directly learns depth from video sequences with view synthesis. In addition, our model accommodates wide field-of-view cameras typically used in colonoscopy and specific challenges such as deformable surfaces, specular lighting, non-Lambertian surfaces, and high occlusion. We performed qualitative analysis on a synthetic data set, a quantitative examination of the colonoscopy training model, and real colonoscopy videos in near real-time.

KEYWORDS

colonoscopy, depth estimation, wide-angle camera, endorobot, navigation

1 Introduction

Colorectal cancer (CRC) causes approximately 900,000 deaths worldwide (Bray et al., 2018) and around 16,800 deaths in the United Kingdom annually. A total of 42,900 new cases were reported in the UK in 2018, making CRC the third most common cancer, accounting for 11% of all new cancer cases. Optical colonoscopy is the gold standard procedure for CRC screening. However, its quality is strongly dependent on the gastroenterologist's skill level (Baxter et al., 2009). Recent studies by le Clercq et al. (2014) have shown that around 60% of CRC cases are related to missed lesion detection from a previous colonoscopy procedure. Chromoendoscopy can increase the visibility of a lesion by sparing and rinsing the colon with a topical dye that increases the colonic wall surface contrast (Durr et al., 2014). However, this procedure requires twice as much time (Pohl et al., 2011), making it undesirable when compared to optical colonoscopy.

The recent introduction of computer-assisted technologies using artificial intelligence for colonoscopy has demonstrated promising results in polyp detection (Lee et al., 2020), size classification (Itoh et al., 2018), and colon coverage (Freedman et al., 2020). The field of robotics has also shown interest in new technologies for colonoscopy (Ciuti et al., 2020), capsules (Formosa et al., 2019), endorobots (Manfredi (2021, 2022)), meshworm robots (Bernth et al., 2017), soft robots (Manfredi et al., 2019), and autonomous robots (Kang et al., 2021). A colonoscope is a long, flexible tubular instrument, generally 12 mm in diameter with a single camera head, lights, irrigation, and an instrument channel port. The miniature camera and the lights allow visual inspection of the colon, and the irrigation port equipment with water helps clean the colon from residual stools. The instrument port enables the passage of surgical tools to remove tissue or polyps for further examination. The size of the colonoscope is restricted by the colon size; thus, the number of sensors to be included, such as three-dimensional (3D) cameras or other sensors, is limited. This has led researchers to investigate 3D mapping from images from a monocular camera inside the colon (Mahmood et al., 2018).

In recent years, monocular depth estimation has been studied with supervised learning (Nadeem and Kaufman, 2016), conditional random fields (Mahmood and Durr, 2018), generative adversarial networks (GANs) (Mahmood et al., 2018), conditional GANs (Rau et al., 2019), and self-supervised learning (Hwang et al., 2021). However, previous works had not considered that the colonoscope camera yields highly distorted images as it uses high field-of-view (FOV) cameras. Cheng et al. (2021) used synthetic data with ground-truth depth to train the baseline supervised depth model used in a self-supervised pipeline and used optical flow from a pre-trained flow network for temporal structural consistency. However, optical flow in a deformable environment like a colon is unreliable. In Bae et al. (2020), the training pipeline follows multiple steps: sparse depth from Structure from Motion (SfM), sparse depth used to train a depth network, pixel-wise embedding, and multi-view reconstruction, which makes it slower and unsuitable to run alongside a robot. The main contributions of the proposed work are:

1. Introduces self-supervised monocular depth estimation methods specifically designed for wide FOV colonoscopy cameras, addressing the challenges posed by the distorted and wide-angle nature of the images.
2. The proposed approach effectively handles low-texture surfaces in view synthesis by incorporating a validity mask, which successfully removes pixels affected by specular lighting, lens artifacts, and regions with zero optical flow, ensuring accurate depth estimation even in challenging areas of the colon.

2 Motivation

Colonoscopy cameras generally have a wide FOV that enables faster diagnosis with complete coverage of the colon wall. The state-of-the-art colonoscopy devices have a 140°–170° FOV camera. As a result, these cameras suffer from high lens distortion, especially radial, as shown in Figure 1, but have not yet been studied sufficiently because most endoscopy depth estimation methods are trained on low FOV images or synthetic images without distortion.

2.1 Low FOV cameras

Low FOV cameras have <120° horizontal FOV lenses as shown in the UCL 2019 data set (Rau et al., 2019). The Brown–Conrady camera model (Kannala and Brandt, 2006) considers these lenses with both radial and tangential distortion. The UCL 2019 data set (Rau et al., 2019) contains synthetic images generated with the ideal pinhole camera approximation. Image distortion is prevalent in real images caused by varying lens magnifications along the increasing angular distance. Due to the imperfect alignment of the lens or sensors, distortion may get decentered. The Brown–Conrady (Kannala and Brandt, 2006) projection function $\pi(P, i)$: $P(x, y, z) \rightarrow p(u, v)$ is defined as

$$
\begin{aligned}
x' &= \frac{\mathrm{x}}{\mathrm{z}} \quad y' = \frac{\mathrm{y}}{\mathrm{z}},\\
x'' &= x'\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) + 2p_1 x' y' + p_2\left(r^2 + 2x'^2\right),\\
y'' &= y'\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) + p_1\left(r^2 + 2y'^2\right) + 2p_2 x' y',\\
u &= x'' f_x + c_x \quad v = y'' f_y + c_y,
\end{aligned}
\tag{1}
$$

where P is a 3D camera point with coordinates (x, y, z), p is a 2D image point with coordinates (u, v), and $r^2 = x^2 + y^2$. Camera parameters i consist of radial distortion coefficients (k_1, k_2, k_3), tangential distortion coefficients (p_1, p_2) of the lens, focal lengths (f_x, f_y), and principal points (c_x, c_y).

2.2 Wide FOV colonoscopy cameras

Omnidirectional camera model (Scaramuzza et al., 2006), unified camera model (UCM) (Geyer and Daniilidis, 2000), extended UCM (eUCM) (Khomutenko et al., 2015), or double sphere (DS) camera model (Usenko et al., 2018) can map wide FOV lenses. The omnidirectional camera model along with the pinhole radial distortion (Fryer and Brown, 1986) and Brown–Conrady (Kannala and Brandt, 2006) camera models falls in the high-order polynomial distortion family. These models require solving the root of a high-order polynomial during unprojection, which is computationally expensive. On the other hand, UCM, eUCM, and DS fall into a unified camera model family that is easy to compute and has a closed-form unprojection function. For briefness, we only describe the DS in this article, but this pipeline can be extended to all unified camera models. DS is modeled with six camera parameters i: $f_x, f_y, c_x, c_y, \alpha$, and ξ.

In the study by Geyer and Daniilidis (2000), it was observed that the UCM effectively represents systems with parabolic, hyperbolic, elliptic, and planar mirrors. Furthermore, the UCM has been successfully employed for cameras equipped with fisheye lenses (Ying and Hu, 2004). However, it is important to acknowledge that the UCM may not provide an ideal match for most fisheye lenses, often necessitating the incorporation of an additional distortion model. The UCM projects a 3D point onto the unit sphere and then onto the image plane of the pinhole camera, which is shifted by ξ from the center of the unit sphere. The eUCM can be considered a generalization of the UCM, where the point is projected onto an ellipsoid symmetric around the z-axis using the coefficient β. Alternatively, the DS (Usenko et al., 2018) camera model is a more suitable choice for cameras with fisheye lenses. It offers a closed-form inverse and avoids computationally expensive trigonometric operations. In DS, a 3D point is projected onto two unit spheres with centers shifted by ξ. Then, the point is projected onto an image plane

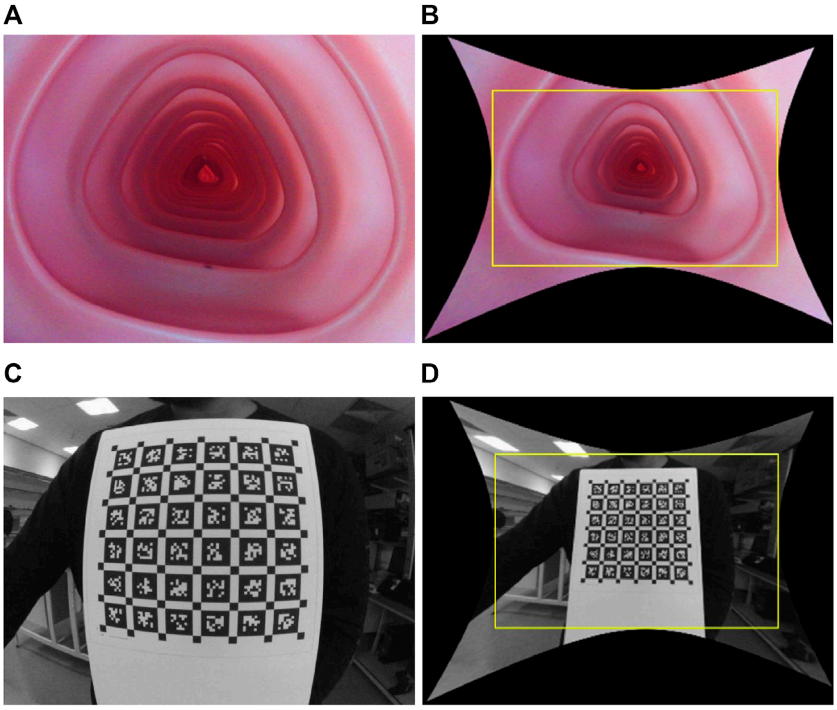

FIGURE 1
Illustration of wide FOV colonoscopy camera and its distortion: raw and rectified images, respectively, when the camera is placed inside the colonoscopy training model **(A)** and **(B)**, and in front of a calibration target **(C)** and **(D)**. The yellow box in **(B)** and **(D)** crops out the black pixels after undistortion, causing a loss of FOV.

using a pinhole shifted by α. It corrects the image distortion that occurs when the camera's image plane 1−α is not perfectly aligned with the object plane. The projection function $\pi(P, i)$: $P(x, y, z) \to p$ (u, v) is given as follows:

$$\pi(P,i) = \begin{bmatrix} f_x \dfrac{x}{x\alpha d_2 + (1-\alpha)(\xi d_1 + z)} \\ f_y \dfrac{y}{\alpha d_2 + (1-\alpha)(\xi d_1 + z)} \end{bmatrix} + \begin{bmatrix} cx \\ cy \end{bmatrix}, \tag{2}$$

where

$$d_1 = \sqrt{x^2 + y^2 + z^2} \quad d_2 = \sqrt{x^2 + y^2 + (\xi d_1 + z)^2}. \tag{3}$$

The unprojection function $\pi^{-1}(p,i)$: $p(u,v) \to P(x,y,z)$ is given as follows:

$$\pi^{-1}(p,i) = \frac{m_z\xi + \sqrt{m_z^2 + (1-\xi^2)r^2}}{m_z^2 + r^2} \begin{bmatrix} m_x \\ m_y \\ m_z \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \xi \end{bmatrix}, \tag{4}$$

where

$$m_x = \frac{u - c_x}{f_x} \quad m_y = \frac{u - c_y}{f_y} \quad r^2 = m_x^2 + m_y^2.$$
$$m_z = \frac{1 - \alpha^2 r^2}{a\sqrt{1 - (1 - (2\alpha - 1)r^2 + 1 - \alpha)}}. \tag{5}$$

3 Methods

In this work, we aim to estimate depth from a colonoscopy image stream via view synthesis. View synthesis enables us to train the depth estimation model in self-supervised mode. The depth network takes one image at a time, source I_s or target I_t image, and predicts the corresponding per-pixel depth maps. Target depth maps D_t from the depth network and $T_{t\to s}$ from pose estimation enable the reconstruction of the target image $\hat{I}_t$ from the source image I_s with geometric projection. The pose network predicts the rigid transformation with six degrees of freedom. With a chosen camera model, the unprojection function π^{-1} can map the target image coordinate $I_t(p)$ to 3D space, forming a 3D point cloud. An overview of the training pipeline is shown in Figure 2. The projection function π maps the 3D points to the target image coordinates. Section 2 describes the camera models for different lenses with varying lens distortions. For camera models that have to find the root of a high-order polynomial, a pre-calculated lookup table can be used for a computationally efficient unprojection operation.

3.1 Generic image synthesis

Irrespective of the camera models shown in Eq **1** or Eq **2-4**, the generic image synthesis block can reconstruct the target image from source images with projection π and unprojection π^{-1} functions, as shown in Figure 3. A point cloud P_t can be generated from the estimated depth D_t for the corresponding input image I_t as follows:

$$P_t = \pi^{-1}(pt, D_t), \tag{6}$$

where p_t is the set of image coordinates and P_t is the 3D point cloud corresponding to the target image I_t. The relative pose $T_{t\to s}$ from the

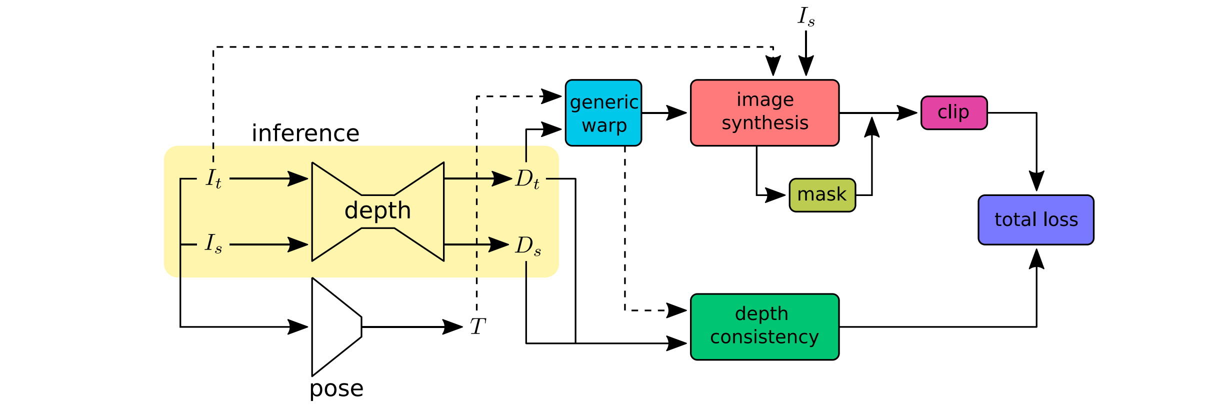

FIGURE 2
Self-supervised depth training pipeline: input images (I_s, I_t) are adjacent frames sampled from a video sequence. The *depth* network predicts independent per-pixel dense depth maps (D_s, D_t) from the source I_s and target I_t images, respectively. The *pose* network takes the concatenated image sequence (I_s, I_t) as input and predicts the rotation R and translation t that define the transformation T between the source and target images. The *warp* function takes the predicted depths and poses to warp the source image coordinates to the target or *vice versa*. The *image synthesis* projects the source image I_s to the target image I_t with the warped image coordinates. The synthesized images are compared with ground-truth images to construct the loss. The image synthesis loss outliers are removed using the *clip* function. Image pixels affected by specular lighting and violation of the Lambertian surface assumption are ignored in loss calculation with the valid *mask*. The *depth consistency* loss ensures that the depth is consistent in the sequence. At inference, only the depth network will be used.

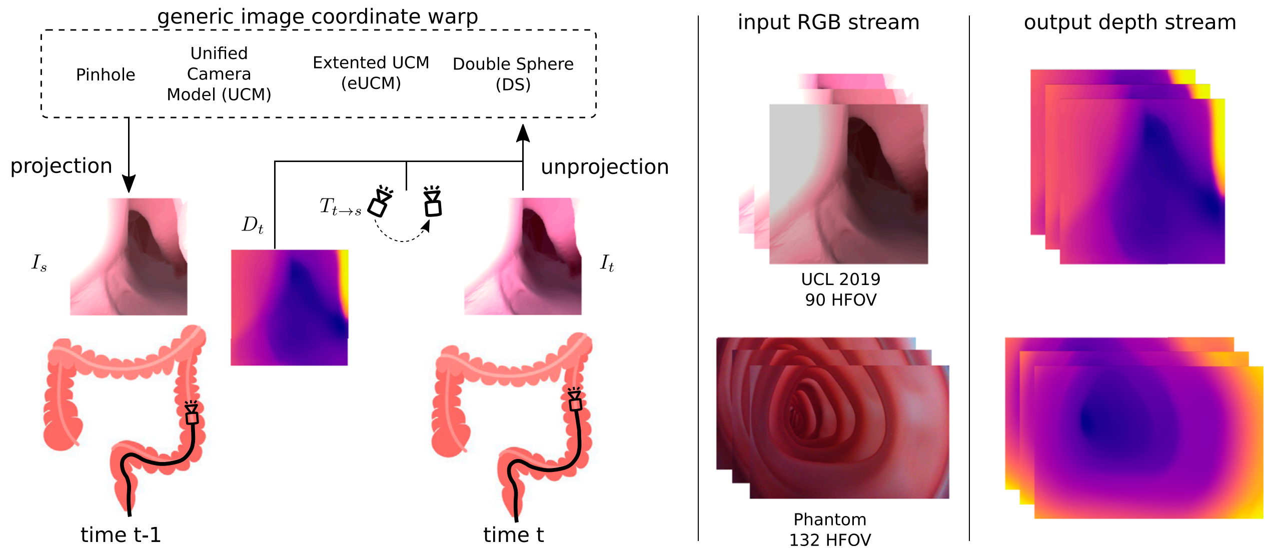

FIGURE 3
Generic image coordinate warp block: source I_s and target I_t images are captured at time $t-1$ and t. The depth D_t predicted by the *depth* network along with the relative pose $T_{t\to s}$ from the source to target estimated by the *pose* network is used to synthesize the source image. The view synthesizing block allows the usage of various camera models that can handle cameras with higher lens distortion. We have demonstrated this use case on the UCL 2019 synthetic data set with a 90° horizontal FOV camera (no lens distortion) and on our colonoscopy training model data set with a 132° horizontal FOV camera (high radial lens distortion).

target to source image is estimated by the pose network. This relative pose can be used to transform the target point cloud into the source point cloud $P_s = T_{t\mapsto s}P_t$. With the projection model π, the 3D point cloud P_s can be mapped to the source image coordinate p_s as follows. The target image can be synthesized by inverse warping the source image coordinates and sampling (Godard et al., 2017) with ζ.

$$p_s = \pi\left(T_{t\to s}\pi^{-1}(p_t, D_t)\right).$$

$$\hat{I}_t^{ij} = \zeta\left(I_s^{i'j'}\right). \tag{7}$$

The projected source image coordinate $p_s(i', j')$ will be used to sample pixel values with the sampling function ζ to generate the target image $\hat{I}_t$. The differentiable spatial transformer network (Jaderberg et al., 2015) is used to synthesize images with bilinear interpolation.

3.2 Image reconstruction loss

The photometric loss L_v between the reference frame and reconstructed frame acts as a self-supervised loss for the depth network. Following the previous works of Zhou et al. (2017), we used structural similarity (SSIM) (Wang et al., 2004) along with the least absolute deviation (L1) for the photometric loss with weight η. This leverages the ability of SSIM, denoted as S, to handle complex illumination changes and add robustness to outliers as follows:

$$L_v^s(I_t\hat{I}_t) = \frac{1}{|V|}\sum_{\mathrm{p}\in\mathrm{V}} \eta \frac{1 - S\left(I_t(p), \hat{I}_t(p)\right)}{2} + (1-\eta)\left|I_t(p) - \hat{I}_t(p)\right|. \tag{8}$$

Here, V denotes valid pixels and $|V|$ denotes its cardinal. Masking out pixels with V that violates non-zero optical flow and the Lambertian surface assumption avoid the propagation of corrupted gradients into the networks. The valid mask is computed as follows:

$$V = M_{ego} \cdot \left(\left|I_t(p) - \hat{I}_t(p)\right| < \left|I_t(p) - I_s(p)\right|\right). \tag{9}$$

Here, M_{ego} is a binary mask that initially marks all mapped pixels as valid. The pixels that appear to be static between the source and target images caused by the absence of ego motion, special lighting, lens artifacts, and low texture are then marked invalid. Parts of the photometric loss that violate the assumptions generate a great loss, yielding a large gradient that potentially worsens the performance. Following the study of Zhou et al. (2018), clipping the loss with θ, which is the p-th percentile of L_v, removes the outlines as $L_v(s_i) = min(s_i,\theta)$, where s_i is the i-th cost in L_v.

3.3 Depth consistency loss

Depth consistency loss L_c (Bian et al., 2021) ensures depth maps D_s and D_t estimated by the depth network with the source and target images I_s and I_t that adhere to the same 3D scene. Minimizing this loss encourages depth consistency not only in the sequence but also to the entire sequence with overlapping image batches. Depth consistency loss can be defined as follows:

$$L_c^s(D_s, D_t) = \frac{1}{|V|}\sum_{\mathrm{p}\in\mathrm{V}} DC^s,$$

$$DC^s = \frac{\left|\hat{D}_s^t(p) - \hat{D}_s(p)\right|}{\hat{D}_s^t(p) - \hat{D}_s(p)}, \tag{10}$$

where $\hat{D}_s^t$ is the projected depth of D_t with pose $T_{t\to s}$ that corresponds to the depth at image I_s, but comparing $\hat{D}_s^t$ with D_s cannot be done as the projection does not lie in the grid of D_s. Thus, $\hat{D}_s^t$ is generated using differentiable bilinear interpolation on D_s and compared with $\hat{D}^t$. The inverse of DC will act as an occlusion mask M_o that can mask out occluded pixels from the source to the target view, mainly around the haustral folds of the colon. The image reconstruction loss L_s is reweighted with M_o mask as follows:

$$L_v^{s'} = \frac{1}{|V|}\sum_{\mathrm{p}\in\mathrm{V}} M_0^s \bullet L_v^s.$$

$$M_0^s = 1 - DC^s. \tag{11}$$

3.4 Edge-guided smoothness loss

The depth maps are regularized with edge-guided smoothness loss as the homogeneous and low-texture regions do not induce information loss in image reconstruction. The smoothness loss is adapted from Wang et al. (2018) as follows:

$$L_e^s(I_t, D_t)\sum\nolimits_{\mathrm{p}\in\mathrm{I}} e^{-\nabla It(p)} \nabla D'(p), \tag{12}$$

where ∇ denotes the first-order gradient and $D_t' = (1/D_t)/mean(1/D_t)$ is the normalized inverse depth. To avoid the optimizer getting trapped in the local minima (Zhou et al., 2017; Godard et al., 2019), the depth estimation and loss are computed at multiple scales. The total loss is averaged over scales (s = 4) and image batches as follows:

$$L = \frac{1}{4}\sum_{s=1}^{4} \sigma_1 L_v^{s'} + \sigma_2 L_c^s + \sigma_3 L_e^s. \tag{13}$$

4 Experiments

In this work, all models are trained in the UCL 2019 synthetic (Rau et al., 2019) and colonoscopy training model data sets and evaluated on a variety of synthetic (Rau et al., 2019; Azagra et al., 2022) and real (Azagra et al., 2022; Ma et al., 2021) colonoscopy data.

4.1 Implementation details

The depth network is inspired by Godard et al. (2019), with ResNet18 (He et al., 2016) as the encoder and multiscale decoder that predicts the depth at four different scales. The pose network shares the same encoder architecture as the depth network, followed by four 2D convolution layers. The models are built using PyTorch (Paszke et al., 2017) with the Adam optimizer (Kingma and Ba, 2014) to minimize the training error. The models are trained with Nvidia A6000 with batch size 25 for 20 epochs at an initial learning rate of 0.0001 and halves after 15 epochs. The depth network sigmoid output d is scaled as $D = 1/(d \cdot 1/(d_{max} - d_{min}) + 1/d_{min})$, where $d_{min} = 0.1$ and $d_{max} = 20.0$ units corresponding to 0.1–20 cm. The network input image size of the UCL 2019 synthetic image (Rau et al., 2019) is 256 × 256 (height × width) and that of the CTM is 512 × 768. These hyperparameters are set empirically as follows: $\eta = 0.85$, $\sigma_1 = 1.0$, $\sigma_2 = 0.5$, and $\sigma_3 = 0.001$. We perform horizontal flipping with 50% of the images in the UCL 2019 data set but not in the CTM data set because of the off-centered principal point of the wide-angle camera. We perform random brightness, contrast, saturation, and hue jitter within ±0.2, ±0.2, ±0.2, and ±0.1 range, respectively. For stability, the first epoch of the training is a warm-up run with minimum reprojection loss (Godard et al., 2019) when trained on the CTM, whereas no warm-up is used for the UCL 2019 synthetic data set. The depth estimation model runs at 10 FPS (frames per second) on the GPU at inference, making it suitable for robotic control.

4.1.1 UCL 2019 synthetic data set

The UCL synthetic data (Rau et al., 2019) are generated based on human CT colonography with manual segmentation and meshing. The unity engine is used to simulate the RGB images with a virtual pinhole camera (90° FOV and two attached light sources) and their corresponding depth maps with a maximum range of 20 cm. The data set contains 15,776 RGB images and ground-truth depth maps, which are split into the 8:1:1 train, validation, and test sets. The data set is recorded as nine subsets with three lighting conditions and one of three different materials. The light sources vary in spot angle, range, color, and intensity, while the materials vary in color, reflectiveness, and smoothness.

4.1.2 Colonoscopy training model data set

Our colon training model data set is recorded with an off-the-shelf full HD camera (1920 × 1,080, 30 Hz, 132° HFOV, 65° VFOV, 158° DFOV, FID 45-20-70, white ring light around camera) inside a plastic phantom used for training medical professionals. The phantom mimics the 1:1 anatomy of the human colon, such as the internal diameter, overall length, and haustral folds, simulating images from an optical colonoscopy. The camera is attached to a motorized shaft that is moved forward and backward while recording videos. The data set contains around 10,000 images split into an 8:2 train and validation set with varying internal and external light settings and does not provide ground-truth depth maps.

4.2 Quantitative study

The unique structural restriction of a real colon precludes the collection of ground-truth depth with a true depth sensor. Thus, quantitative depth evaluation can only be conducted on synthetic data such as UCL 2019, even though the current synthetic data generation is not realistic enough to mimic real colon surface properties. The UCL synthetic data contains some texture, such as blood veins and haustral folds. The proposed depth estimation is assessed on the UCL 2019 test data set with evaluation metrics by Eigen et al. (2014) and compared with the other state-of-the-art methods like SfMLearner (Zhou et al., 2017), Monodepth1 (Godard et al., 2017), DDVO (Wang et al., 2018), Monodepth2 (Godard et al., 2019), HR Depth (Lyu et al., 2021), AF-SfMLearner (Shao et al., 2021), and AF-SfMLearner2 (Shao et al., 2022) in Table 1.

4.3 Qualitative analysis

Quantitative depth evaluation of real colonoscopy data like EndoMapper (Azagra et al., 2022) or LDPolypVideo (Ma et al., 2021) is challenging. Much of the real characteristics of the colon, like reflectance of the surface and light diffusion of different tissue types, are hard to mimic in a simulator. In addition, a real colon is not rigid; the deformable property of the colon wall makes the view synthesis even more problematic. It was observed that the depth network was leaving holes in the depth maps when trained on the low-textured CTM with just image reconstruction loss. Low-textured regions often occur in the real colon wall, and handling these regions is critical for data-driven methods that rely on view synthesis. The valid mask described in Section 3.2 works well for removing pixels affected by specular lighting, lens artifacts, and zero optical flow regions. The data collection of the CTM was limited to the linear motion of the steady-speed motorized shaft inside the phantom on which the wide-angle camera was fixed. This limits the prediction of the depth network when introduced to real colon images with sharp turns. The depth consistency loss described in Section 3.3 helps the network to learn geometric consistency and avoid depth map flickering between frames. The depth network trained on the CTM learns to directly predict depth from highly distorted images with the help of the generic image synthesizer briefed in Section 3.1 that can handle a variety of camera models. The model trained with a pinhole camera was tested against

TABLE 1 Quantitative study on proposed depth model on UCL 2019 synthetic data. Best values are in bold, and the second best values are underlined. *Abs Rel*, *Sq Rel*, *RMSE*, and *RMSE log* are in mm, and $\delta < 1.25^1$, $\delta < 1.25^2$, and $\delta < 1.25^3$ are in percentage.

Model	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25^1$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
SfMLearner (Zhou et al., 2017)	0.451	0.711	1.768	0.537	0.366	0.618	0.785
Monodepth1 (Godard et al., 2017)	0.444	0.752	1.755	0.531	0.371	0.625	0.790
DDVO (Wang et al., 2018)	0.452	0.772	1.768	0.538	0.366	0.618	0.785
Monodepth2 (Godard et al., 2019)	0.446	0.756	1.757	0.532	0.360	0.624	0.789
HR Depth (Lyu et al., 2021)	0.448	0.762	1.762	0.534	0.369	0.621	0.788
AF-SfMLearner (Shao et al., 2021)	0.387	0.618	1.648	0.480	0.420	0.686	0.831
AF-SfMLearner2 (Shao et al., 2022)	0.352	0.545	1.581	0.448	0.445	0.712	0.852
Our	**0.141**	**0.115**	**0.669**	**0.179**	**0.828**	**0.959**	**0.985**

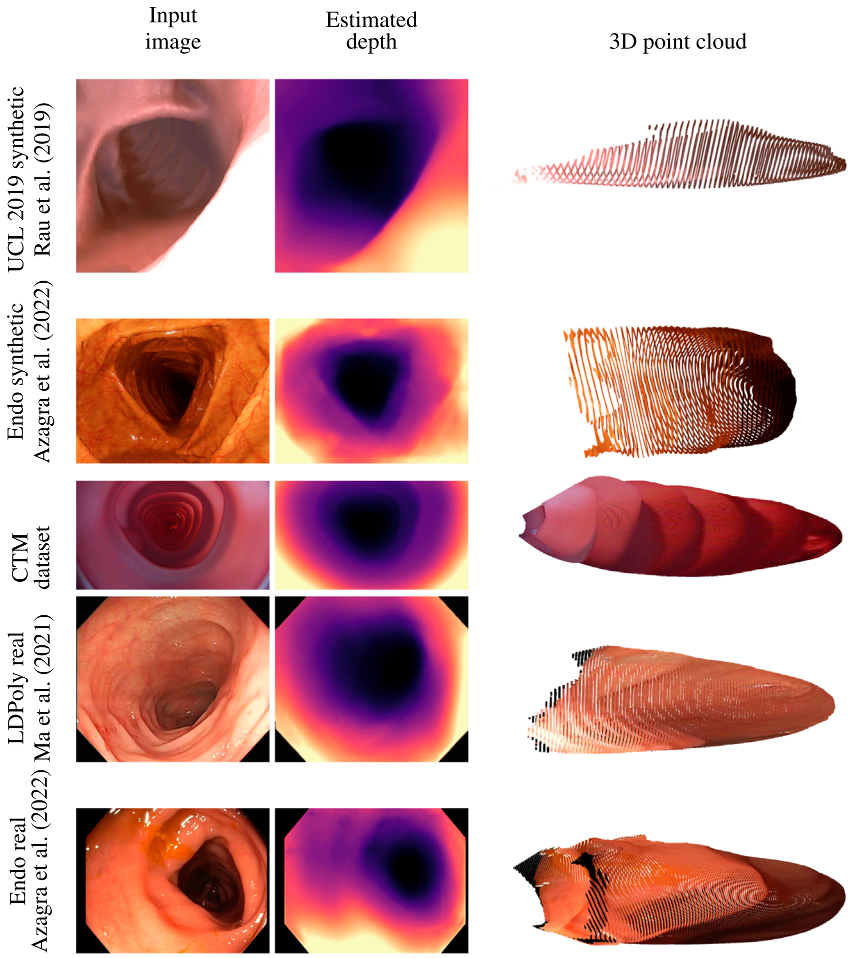
Input image
Estimated depth
3D point cloud
UCL 2019 synthetic Rau et al. (2019)
Endo synthetic Azagra et al. (2022)
CTM dataset
LDPoly real Ma et al. (2021)
Endo real Azagra et al. (2022)

FIGURE 4
Qualitative analysis on synthetic data sets like UCL 2019 synthetic and EndoMapper synthetic data sets, CTM data set, and real colonoscopy images like LDPolypVideo and EndoMapper data set (top to bottom) (Azagra et al., 2022; Ma et al., 2021).

TABLE 2 Ablation study on UCL 2019 synthetic data. *View*, *Valid*, *Consistency*, and *Clip* represent image view synthesis, valid mask, depth consistency, and loss clipping, respectively. Best values are in bold and second best values are underlined. *Abs Rel*, *Sq Rel*, *RMSE*, and *RMSE log* are in mm, and $\delta < 1.25^1$, $\delta < 1.25^2$, and $\delta < 1.25^3$ are in percentage.

View	Valid	Consistency	Clip	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25^1$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
✓	✗	✗	✗	0.195	0.184	0.843	0.236	0.748	0.909	0.961
✓	✓	✗	✗	0.180	0.150	0.734	0.216	0.774	0.928	0.968
✓	✗	✗	✓	0.157	0.139	0.703	0.195	**0.803**	0.945	0.981
✓	✓	✗	✓	**0.141**	**0.115**	**0.669**	**0.179**	0.828	**0.959**	**0.985**
✓	✗	✓	✗	0.179	0.144	0.725	0.215	0.779	0.929	0.970
✓	✓	✓	✗	0.171	0.139	0.729	0.209	0.791	0.933	0.971
✓	✓	✓	✓	0.176	0.140	0.699	0.211	0.788	0.931	0.970

an unseen EndoMapper synthetic data set (Azagra et al., 2022). The model trained on the wide-angle camera was tested against real colonoscopy data sets like EndoMapper (Azagra et al., 2022) and LDPolypVideo (Ma et al., 2021). It is to be noted that the model trained on a wide-angle camera performed better in real colonoscopy data sets that used wide FOV cameras than the model trained on synthetic data, as shown in Figure 4. The model trained on rectified or perfect pinhole camera images, as in UCL 2019, is suboptimal to the target colonoscopy camera, which is distorted and wide angled. If trained on raw images like in the CTM, the network learns distortion as part of the transfer function, and it is only weakly encoded and thus expected to be more robust.

4.4 Ablation study

Table 2 demonstrates the impact of each loss when trained with UCL synthetic data. The depth evaluation reflects the benefit of the valid mask in Section 3.2, which ignores pixels that do not contribute to the photometric loss. It is also found that clipping off high photometric loss improves model performance. Even though depth consistency in Section 3.3 leads to consistent depth in the sequence, it did not improve quantitative results but helped qualitatively on the CTM data set.

5 Conclusion

This work presents a self-supervised depth estimation model for wide-angle colonoscopy cameras. To the best of our knowledge, this is the first time data from a wide-angle colonoscopy camera is used to estimate depth from a video sequence. Most of the previous works had assumed a pinhole model for the colonoscopy camera, but real colonoscopy cameras have a wide FOV lens. Our network predicts depth directly on highly distorted raw images typical for such real cameras. The pipeline also focuses on handling texture regions in the images that generate low photometric loss and geometrically consistent depth estimation in the sequence. Our methods fill a gap in modeling wide FOV cameras and low-texture regions in colonoscopy imaging. We also achieved near-real-time prediction with the depth estimation models that allow us to operate alongside a colonoscopy robot.

A limitation of our work is that the UCL 2019 synthetic data set does not fully capture the complex surface properties and characteristics of real colon tissue. The absence of ground-truth depth maps for the CTM data set further limits the ability to quantitatively assess performance on this data set. Further exploration and evaluation of diverse and realistic data sets can enhance the generalizability and reliability of the proposed depth estimation method.

Data availability statement

The data sets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article.

Author contributions

AM contributed to the design and implementation of the code, experimental setup, measurement, and data analysis and drafted the first version of the manuscript. LiM, ET, and LdM contributed to the scientific contents, coordination, and manuscript writing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) (grant no. EP/W00433X/1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, editors, and reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Azagra, P., Sostres, C., Ferrandez, Á., Riazuelo, L., Tomasini, C., Barbed, O. L., et al. (2022). *EndoMapper dataset of complete calibrated endoscopy. ArXiv. /abs/2204.14240.*

Bae, G., Budvytis, I., Yeung, C.-K., and Cipolla, R. (2020). "Deep multi-view stereo for dense 3d reconstruction from monocular endoscopic video," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 29 September 2020 (Springer), 774–783. doi:10.1007/978-3-030-59716-0_74

Baxter, N. N., Goldwasser, M. A., Paszat, L. F., Saskin, R., Urbach, D. R., and Rabeneck, L. (2009). Association of colonoscopy and death from colorectal cancer. *Ann. Intern. Med.* 150, 1–8. doi:10.7326/0003-4819-150-1-200901060-00306

Bernth, J. E., Arezzo, A., and Liu, H. (2017). A novel robotic meshworm with segment-bending anchoring for colonoscopy. *IEEE Robotics Automation Lett.* 2, 1718–1724. doi:10.1109/lra.2017.2678540

Bian, J.-W., Zhan, H., Wang, N., Li, Z., Zhang, L., Shen, C., et al. (2021). Unsupervised scale-consistent depth learning from video. *Int. J. Comput. Vis.* 129, 2548–2564. doi:10.1007/s11263-021-01484-6

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a cancer J. Clin.* 68, 394–424. doi:10.3322/caac.21492

Cheng, K., Ma, Y., Sun, B., Li, Y., and Chen, X. (2021). "Depth estimation for colonoscopy images with self-supervised learning from videos," in International conference on medical image computing and computer-assisted intervention, 21 September 2021 (Springer), 119–128. doi:10.1007/978-3-030-87231-1_12

Ciuti, G., Skonieczna-Zydecka, K., Marlicz, W., Iacovacci, V., Liu, H., Stoyanov, D., et al. (2020). Frontiers of robotic colonoscopy: A comprehensive review of robotic colonoscopes and technologies. *J. Clin. Med.* 9, 1648. doi:10.3390/jcm9061648

Durr, N. J., Gonza´lez, G., and Parot, V. (2014). *3d imaging techniques for improved colonoscopy.*

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Adv. neural Inf. Process. Syst.* 27. doi:10.5555/2969033.2969091

Formosa, G. A., Prendergast, J. M., Edmundowicz, S. A., and Rentschler, M. E. (2019). Novel optimization-based design and surgical evaluation of a treaded robotic capsule colonoscope. *IEEE Trans. Robotics* 36, 545–552. doi:10.1109/tro.2019.2949466

Freedman, D., Blau, Y., Katzir, L., Aides, A., Shimshoni, I., Veikherman, D., et al. (2020). Detecting deficient coverage in colonoscopies. *IEEE Trans. Med. Imaging* 39, 3451–3462. doi:10.1109/tmi.2020.2994221

Fryer, J. G., and Brown, D. C. (1986). Lens distortion for close-range photogrammetry. *Photogrammetric Eng. remote Sens.* 52, 51–58.

Geyer, C., and Daniilidis, K. (2000). "A unifying theory for central panoramic systems and practical implications," in Computer vision — ECCV 2000. ECCV 2000. Lecture notes in computer science (Berlin, Heidelberg: Springer), Vol. 1843. doi:10.1007/3-540-45053-X_29

Godard, C., Aodha, O. M., and Brostow, G. J. (2017). "Unsupervised monocular depth estimation with left-right consistency," in 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, 6602–6611. doi:10.1109/CVPR.2017.699

Godard, C., Aodha, O. M., Firman, M., and Brostow, G. (2019). "Digging into self-supervised monocular depth estimation," in 2019 IEEE/CVF international conference on computer vision (ICCV), Seoul, Korea (South), 3827–3837. doi:10.1109/ICCV.2019.00393

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, 770–778. doi:10.1109/CVPR.2016.90

Hwang, S.-J., Park, S.-J., Kim, G.-M., and Baek, J.-H. (2021). Unsupervised monocular depth estimation for colonoscope system using feedback network. *Sensors* 21, 2691. doi:10.3390/s21082691

Itoh, H., Roth, H. R., Lu, L., Oda, M., Misawa, M., Mori, Y., et al. (2018). "Towards automated colonoscopy diagnosis: Binary polyp size estimation via unsupervised depth learning," in International conference on medical image computing and computer-assisted intervention, 26 September 2018 (Springer), 611–619. doi:10.1007/978-3-030-00934-2_68

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. *Adv. neural Inf. Process. Syst.* 28.

Kang, M., Joe, S., An, T., Jang, H., and Kim, B. (2021). A novel robotic colonoscopy system integrating feeding and steering mechanisms with self-propelled paddling locomotion: A pilot study. *Mechatronics* 73, 102478. doi:10.1016/j.mechatronics.2020.102478

Kannala, J., and Brandt, S. S. (2006). A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Trans. pattern analysis Mach. Intell.* 28, 1335–1340. doi:10.1109/tpami.2006.153

Khomutenko, B., Garcia, G., and Martinet, P. (2015). An enhanced unified camera model. *IEEE Robotics Automation Lett.* 1, 137–144. doi:10.1109/lra.2015.2502921

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

le Clercq, C. M., Bouwens, M. W., Rondagh, E. J., Bakker, C. M., Keulen, E. T., de Ridder, R. J., et al. (2014). Postcolonoscopy colorectal cancers are preventable: A population-based study. *Gut* 63, 957–963. doi:10.1136/gutjnl-2013-304880

Lee, J. Y., Jeong, J., Song, E. M., Ha, C., Lee, H. J., Koo, J. E., et al. (2020). Real-time detection of colon polyps during colonoscopy using deep learning: Systematic validation with four independent datasets. *Sci. Rep.* 10, 8379–9. doi:10.1038/s41598-020-65387-1

Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., et al. (2021). "Hr-depth: High resolution self-supervised monocular depth estimation," in Proceedings of the AAAI Conference on Artificial Intelligence, 2294–2301. doi:10.1609/aaai.v35i3.1632935

Ma, Y., Chen, X., Cheng, K., Li, Y., and Sun, B. (2021). "Ldpolypvideo benchmark: A large-scale colonoscopy video dataset of diverse polyps," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 21 September 2021 (Springer), 387–396. doi:10.1007/978-3-030-87240-3_37

Mahmood, F., Chen, R., and Durr, N. J. (2018). Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Trans. Med. imaging* 37, 2572–2581. doi:10.1109/tmi.2018.2842767

Mahmood, F., and Durr, N. J. (2018). Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Med. image Anal.* 48, 230–243. doi:10.1016/j.media.2018.06.005

Manfredi, L., Capoccia, E., Ciuti, G., and Cuschieri, A. (2019). A soft pneumatic inchworm double balloon (spid) for colonoscopy. *Sci. Rep.* 9, 11109–9. doi:10.1038/s41598-019-47320-3

Manfredi, L. (2022). *Endorobotics: Design, R&D and future trends*. Academic Press.

Manfredi, L. (2021). Endorobots for colonoscopy: Design challenges and available technologies. *Front. Robotics AI* 8, 705454. doi:10.3389/frobt.2021.705454

Nadeem, S., and Kaufman, A. (2016). *Depth reconstruction and computer-aided polyp detection in optical colonoscopy video frames. arXiv preprint arXiv:1609.01329.*

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic differentiation in pytorch," in *NIPS-W*.

Pohl, J., Schneider, A., Vogell, H., Mayer, G., Kaiser, G., and Ell, C. (2011). Pancolonic chromoendoscopy with indigo carmine versus standard colonoscopy for detection of neoplastic lesions: A randomised two-centre trial. *Gut* 60, 485–490. doi:10.1136/gut.2010.229534

Rau, A., Edwards, P., Ahmad, O. F., Riordan, P., Janatka, M., Lovat, L. B., et al. (2019). Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *Int. J. Comput. assisted radiology Surg.* 14, 1167–1176. doi:10.1007/s11548-019-01962-w

Scaramuzza, D., Martinelli, A., and Siegwart, R. (2006). "A flexible technique for accurate omnidirectional camera calibration and structure from motion," in Fourth IEEE International Conference on Computer Vision Systems (ICVS'06), New York, NY, 45–45. doi:10.1109/ICVS.2006.3

Shao, S., Pei, Z., Chen, W., Zhang, B., Wu, X., Sun, D., et al. (2021). "Self-supervised learning for monocular depth estimation on minimally invasive surgery scenes," in 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi an, China, 30 May 2021 - 05 June 2021 (IEEE), 7159–7165. doi:10.1109/ICRA48506.2021.9561508

Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., et al. (2022). Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. *Med. image Anal.* 77, 102338. doi:10.1016/j.media.2021.102338

Usenko, V., Demmel, N., and Cremers, D. (2018). "The double sphere camera model," in 2018 International Conference on 3D Vision (3DV) (IEEE), 552–560.

Wang, C., Buenaposada, J. M., Zhu, R., and Lucey, S. (2018). "Learning depth from monocular videos using direct methods," in Proceedings of the IEEE conference on computer vision and pattern recognition. –2030.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Trans. image Process.* 13, 600–612. doi:10.1109/tip.2003.819861

Ying, X., and Hu, Z. (2004). Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model. In Proceedings, Part I 8 Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. (Springer), 442–455.

Zhou, L., Ye, J., Abello, M., Wang, S., and Kaess, M. (2018). *Unsupervised learning of monocular depth estimation with bundle adjustment, super-resolution and clip loss. arXiv preprint arXiv:1812.03368.*

Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). "Unsupervised learning of depth and ego-motion from video," in Proceedings of the IEEE conference on computer vision and pattern recognition, 1851–1858.