Unsupervised Photometric-Consistent Depth Estimation from Endoscopic Monocular Video

Shijie Li[1*], Weijun Lin[1*], Qingyuan Xiang[1], Yunbin Tu[2], Shitan Asu[1], Zheng Li [1†]

[1]College of Computer Science, Sichuan University, Chengdu, China

[2] School of Computer Science and Technology, University of Chinese Academy of Sciences Beijing, China

lizheng@scu.edu.cn

Abstract

Recent advancements in unsupervised monocular depth estimation typically rely on an assumption that image photometry remains consistent across consecutive frames. However, this assumption often fails in endoscopic scenes due to: 1) local photometric inconsistency caused by specular reflections creating highlights; and 2) global photometric inconsistency resulting from the simultaneous movement of the light source and the camera. Since unsupervised depth estimation methods rely on appearance discrepancies between frames as a supervisory signal, these photometric inconsistencies inevitably deteriorate loss function calculation. In this paper, our goal is to obtain a strong and reliable supervisory signal for achieving photometric-consistent depth estimation. To this end, for local photometric inconsistency, we utilize the specular reflection model to introduce a Highlight Loss for handling the estimation of highlight regions. For global photometric inconsistency, we design a Photometric Match module, which utilizes the spotlight illumination model to derive an analytical expression, achieving photometric alignment across different frames. Unlike previous works that introduce additional optical flow or networks, our method is simpler and more efficient. Extensive experiments demonstrate our method achieves the state-of-the-art results on C3VD, SCARED and SERV-CT datasets.

Code — https://github.com/DpEstimation/PC-Depth

Introduction

Accurate depth estimation is essential for reconstructing 3D structures from monocular endoscopic videos, which significantly advances minimally invasive surgical navigation (Fitzpatrick 2010; Edwards et al. 2021; Taylor et al. 2016). However, compared to conventional images, the depth prediction in endoscopic scenes often have more complex lighting environments.

Recently, deep learning-based methods (Eigen, Puhrsch, and Fergus 2014; Liu et al. 2015) have made significant strides, which employ Convolutional Neural Networks (CNNs) to predict depth maps from conventional monocular video. In particular, unsupervised methods (Zou, Luo, and Huang 2018; Mahjourian, Wicke, and Angelova 2018a; Bian et al. 2019) leverage CNN-based depth and ego-motion networks, eliminating the need for ground-truth data. The core idea of these methods is to warp the source frame to target frame for generating view synthesis using the predicted depth and camera pose. Subsequently, the appearance discrepancy between the target frame and the synthesized frame serves as a supervisory signal throughout the training phase.

Unfortunately, these methods are not suitable for endoscopic scenarios because they fail to satisfy a critical assumption: Photometric consistency across adjacent frames (Horn and Schunck 1981). This inconsistency stems from two major sources: (1) **Local photometric inconsistency**: Specular reflections create highlights, which cause the depth network to mistakenly treat these highlights as distinct objects, and lead to incorrect depth estimations. (2) **Global photometric inconsistency**: The joint movement of the light source and the camera alters the photometric of same target in different frames, thereby undermining the accuracy of the loss calculation. Consequently, fluctuations in photometric values impede the utilization of appearance differences as a supervisory signal, resulting in ambiguous supervision in endoscopic scenes.

In this paper, we propose Unsupervised Photometric-Consistent Depth Estimation Network (PC-Depth) to tackle these challenges. Our key insight is that, by leveraging a fundamental property of endoscopic imagery, we can establish appropriate and reliable constraints to achieve photometric-consistent depth prediction. Architecture-wise, given two frames (source frame and target frame), we estimate the target frame's depth map and their relative camera pose using depth and pose networks. According to the reflection model, highlights occur when the reflected light coincides with the observation direction (the light source position can be approximated as coinciding with the camera position in endoscopes (Arnold et al. 2010)). To exploit this property and achieve local photometric consistency, we propose a Highlight Loss to constrain the depth estimation within the depth network for highlighted regions. Specifically, the angle between the reflected light and the observation direction is twice the angle between the normal vector and the observation direction. We first derive the normal vectors of the highlight regions from the depth map. Then, we minimize the angle between the normal vector and the observation di-

[*]These authors contributed equally.

[†]Corresponding author

rection, thereby refining the predicted depths.

For achieving global photometric consistency, our goal is to adjust the photometric values of synthetic frame to align with target frame. To this end, we introduce the Photometric Match (PM) module, which comprises two key components: Photometric Assessment and Aligned Wrap. The PM module first establishes an analytical expression to compute the photometric ratio between the source frame and the target frame based on the spotlight illumination model in endoscopy. Subsequently, it utilizes the depth map and ego-motion to wrap the source frame, generating a synthetic frame. Finally, the computed photometric ratio is applied to adjust the photometric values of the synthetic frame. Consequently, this module ensures that the appearance differences between the synthetic frame and the target frame serve as a reliable supervisory signal. **To the best of our knowledge**, this is the first work to utilize the principles of endoscopic imaging to accomplish photometric-consistent depth estimation from monocular endoscopic video.

Our contributions can be summarized as follows:

- We conduct an in-depth analysis of the inherent limitations of traditional unsupervised methods for predicting depth maps in endoscopic scenarios, primarily attributed to local and global photometric inconsistencies.
- We propose a novel unsupervised depth estimation network that provides a strong and reliable supervisory signal to enforce photometric consistency in depth estimation. Unlike previous work, our method does not introduce additional optical flow or auxiliary networks.
- Extensive experiments and analysis demonstrate the effectiveness of our designed components in improving the depth estimation accuracy. Our PC-Depth achieves state-of-the-art performance on the three benchmarks. Further, our components can be integrated into existing baselines, enhancing their performance in endoscopic scenarios.

Related Work

Compared to conventional images (Li et al. 2024a,b; Tu et al. 2024a,b), there are the absence of rich texture information and the influence of complex lighting conditions in endoscopic images, so traditional algorithms (Ren et al. 2017; Recasens et al. 2021), such as simultaneous localization and mapping (SLAM) (Chen et al. 2018), encounter significant challenges in accurately estimating depth.

Consequently, fully supervised convolutional neural network (CNN) have been developed to predict depth maps from monocular videos (Xu et al. 2017; Cao, Wu, and Shen 2017). For instance, Eigen et al .(Eigen, Puhrsch, and Fergus 2014) designed a coarse-to-fine network that predicts depth from a single view, using ground truth depths obtained from range sensors as the supervisory signal. Inspired by this approach, numerous subsequent studies have significantly enhanced the accuracy of depth estimation in various ways (Yin et al. 2019; He, Wang, and Hu 2018; Xu et al. 2018; Zhuang et al. 2022).

Despite fully supervised depth estimation methods have shown significant progress, acquiring ground truth data remains expensive. To eliminate the need for costly depth annotations, Zhou et al. (Zhao et al. 2017) developed an unsupervised framework that jointly trained a depth network and a pose network using unlabeled video sequences. This framework utilized the differences between the target frame and the synthetic frame as the supervisory signal. Inspired by this work, many subsequent studies extended it by introducing additional geometric priors (Yang et al. 2018; Mahjourian, Wicke, and Angelova 2018b; Chen, Schmid, and Sminchisescu 2019; Bian et al. 2019; Yan et al. 2023; Wang et al. 2024; Zhao et al. 2022). For instance, Bian et al. (Bian et al. 2019) proposed a geometry consistency loss for scale-consistent predictions and an induced self-discovered mask for handling moving objects and occlusions. Recent work, such as DepthAnything (Yang et al. 2024a), leveraged large-scale and diverse training data to develop a robust foundational model for depth estimation.

However, these methods may not be applicable to endoscopic scenarios due to photometric inconsistency. To mitigate this limitation, Liu et al. (Liu et al. 2019) utilized Structure from Motion (SfM) techniques to obtain sparse depth maps as supervisory signals. This approach addressed the issue but proved to be cumbersome and time-consuming. Li et al. (Li et al. 2020) incorporated the peak signal-to-noise ratio (PSNR) as an additional optimization objective during the training phase. Endo-SfM (Ozyoruk et al. 2021) modeled the photometric variations between endoscopic images as affine transformations to address this issue. However, actual variations in endoscopic lighting were more complex and could not be accurately represented solely by affine transformations. Recently, AF-SfM (Shao et al. 2022) employed the first-order Taylor expansion method to express photometric transformation as a function of optical flow and the source image. Additionally, it introduced an extra network to predict photometric changes. Essentially, the strategy employed by AF-SfM relies on linear approximations of instantaneous photometric transformations. This coarse approximation cannot adequately retain the power when photometric variation is relatively drastic.

Method

Overview

As shown in Figure 1, given source and target frames (I_s, I_t) sampled from an unlabeled video, we first estimate the target frame's depth map D_t using the depth network and then predict the relative 6D camera pose $T_{t\to s}$ between them using the pose network. The Highlight Loss $\mathcal{L}_H$ is used to constrain the depth network's estimation for highlight regions. Subsequently, the Photometric Match (PM) generates a synthetic frame $\tilde{I}_t^s$, which matches the photometric values of the target frame. The discrepancy between I_t and $\tilde{I}_t^s$ is used as the Photometric Loss $\mathcal{L}_P$. Additionally, inspired by these works (Baker and Matthews 2004; Bian et al. 2019), we introduce Smoothness Loss $\mathcal{L}_S$ to handle areas with missing textures and Geometry Consistency Loss $\mathcal{L}_{GC}$ to ensure scale consistency. Note that highlights within I_t are removed when calculating the Smoothness Loss. In short, our loss function can be formulated as follows:

$$\mathcal{L} = \alpha\mathcal{L}_P + \beta\mathcal{L}_H + \gamma\mathcal{L}_S + \omega\mathcal{L}_{GC} \tag{1}$$

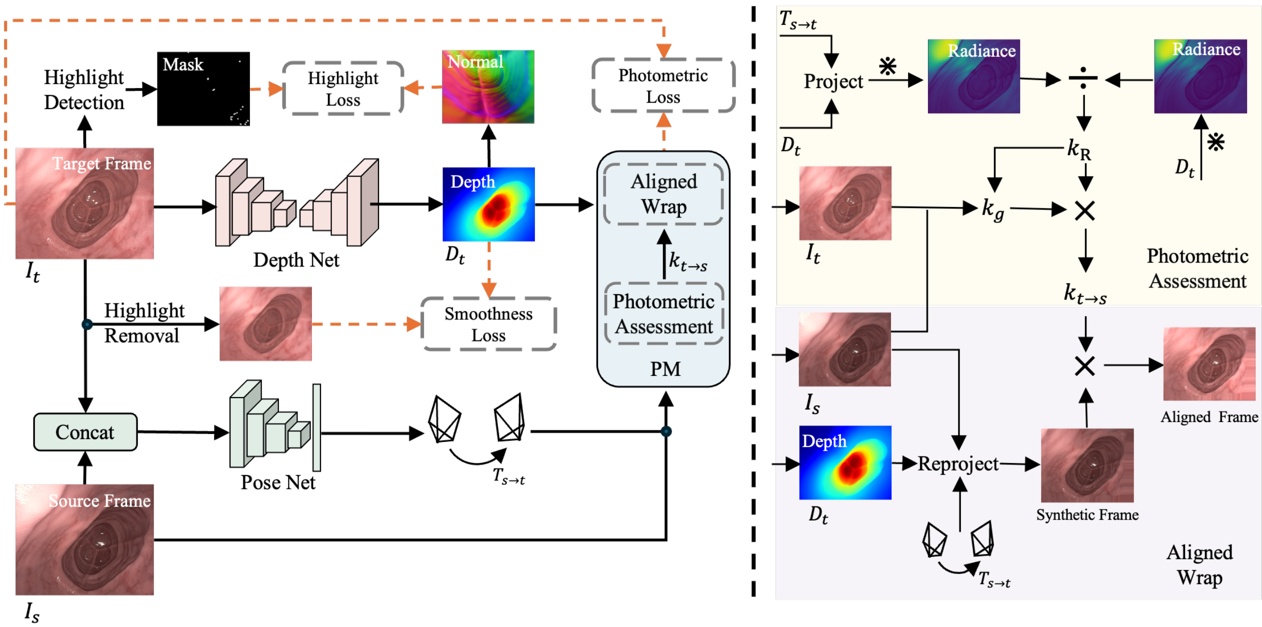

Figure 1: Network Architecture. Left is the overview of our PC-Depth. For clarity, the Geometry Consistency Loss is not shown in this figure. Right is the detail of Photometric Match (PM) module which contains two key component: Photometric Assessment and Aligned Wrap. $*$ denotes radiance calculation.

Subsequent sections will elaborate on the following formulations: (1) Highlight Loss, (2) Photometric Match, (3) Photometric Loss and (4) Smoothness Loss and Geometry Consistency Loss.

Highlight Loss

To constrain local photometric inconsistency caused by highlight, we first employ the threshold-based method (Arnold et al. 2010) to detect the highlight mask M_h. Subsequently, we employ specular reflection model to help establish the relationship between highlights and depth, as illustrated in Figure 2 (a). Specifically, the direction of light reflection can be computed from the direction of light incidence and the normal vector to the surface. The equation is formulated as follows:

$$\boldsymbol{v}_i = \boldsymbol{l}_i - 2(\boldsymbol{l}_i \cdot \boldsymbol{n}_i) \cdot \boldsymbol{n}_i \tag{2}$$

where $\boldsymbol{n}_i$ represents the normal vector of the object's surface, derived from the depth map; $\boldsymbol{s}_i$ is the unit vector in the viewing direction; $\boldsymbol{l}_i$ represents the unit vector in the direction of light incidence; $\boldsymbol{v}_i$ is the unit vector in the direction of light reflection. Since center of the light source coinciding with the optical center of the camera, the viewing direction $\boldsymbol{s}_i$ is exactly opposite to the direction of light incidence $\boldsymbol{l}_i$: $\boldsymbol{l}_i = -\boldsymbol{s}_i$. As a result, the intensity of specular reflection is inversely proportional to the angle θ_i between the viewing direction $\boldsymbol{s_i}$ and the reflection direction $\boldsymbol{v_i}$. In the endoscopic reflection model, the angle can be computed as follows:

$$\theta_i = \arccos(\boldsymbol{s_i} \cdot \boldsymbol{n_i}) \tag{3}$$

where the angle θ_i between the unit vector of viewing direction $\boldsymbol{s}_i$ and the unit vector of reflection direction $\boldsymbol{v}_i$ is twice the angle between the viewing direction $\boldsymbol{s}_i$ and the normal $\boldsymbol{n}_i$.

Assuming the optical center of the camera and the center of the light source are located at the origin of the camera's coordinate system, the unit vector in the viewing direction $\boldsymbol{s}_i$, integrated with the camera imaging model, can be described by the following equation:

$$\begin{aligned} \boldsymbol{p_i} &= D_t(\boldsymbol{p}_i^{uv})\mathbf{K}^{-1}\boldsymbol{p}_i^{uv} \\ \boldsymbol{s}_i &= \mathbf{0} - \boldsymbol{p}_i = -D_t(\boldsymbol{p}_i^{uv})\mathbf{K}^{-1}\boldsymbol{p}_i^{uv} \end{aligned} \tag{4}$$

where $\boldsymbol{p}_i$ represents the 3D spatial point that corresponds to the point $\boldsymbol{p}_i^{uv}$ in the pixel coordinate system, and $\boldsymbol{p}_i$ is under its camera coordinate system; D is the depth map; $\mathbf{K}$ represents the camera intrinsic matrix. Combining Equation 4, the specular reflection intensity in endoscopic images adheres to the following relationship:

$$\text{Specular reflection intensity} \propto (-D_t(\boldsymbol{p}_i^{uv}) \cdot \mathbf{K}^{-1}\boldsymbol{p}_i^{uv}) \cdot \boldsymbol{n}_i \tag{5}$$

where $\propto$ denotes a proportional relationship. Based on this formula, we can define the Highlight Loss $\mathcal{L}_H$ as follows:

$$\begin{aligned} \mathcal{L}_H &= \frac{1}{|M_h|} \sum_{\boldsymbol{p}_i \in M_h} (1 - \boldsymbol{s}_i \cdot \boldsymbol{n}_i)^2 \\ &= \frac{1}{|M_h|} \sum_{\boldsymbol{p}_i \in M_h} \left(1 - (-D_t(\boldsymbol{p}_i) \cdot \mathbf{K}^{-1}\boldsymbol{p}_i) \cdot \boldsymbol{n}_i\right)^2 \end{aligned} \tag{6}$$

where M_h denote the highlight area in the target view, and $|M_h|$ defines the number of points in M_h. This formulation underscores the role of specular reflection intensity as a critical factor in the evaluation of endoscopic image quality.

Photometric Match

To ensure global photometric consistency, we introduce a Photometric Match module, comprising two primary modules: (1) Photometric Assessment and (2) Aligned Warp.

Photometric Assessment. This component aims to leverages the spotlight illumination model (Modrzejewski et al. 2020) in endoscopy to derive the inter-frame photometric ratio. As shown in Figure 2 (b), the position of the spotlight is $\boldsymbol{p}_l$, the unit direction vector of the spotlight is denoted as $\boldsymbol{l}$, we write the radiance of position $\boldsymbol{p}_i$ as:

$$\sigma_{SLS}(\boldsymbol{p}_i, \psi_i) = \frac{\sigma_0}{||\boldsymbol{p}_i - \boldsymbol{p}_l||^2} R(\psi_i) \tag{7}$$

$$R(\psi_i) = e^{-\mu(1-\cos(\psi_i))} \tag{8}$$

where σ_0 is the maximum radiance and $R(\psi_i)$ is the radial attenuation controlled by a spread factor μ; ψ_i denotes off-axis angle. Assuming all surfaces satisfy Lambertian reflection, for each pixel, the rendering equation can be written as follows:

$$\hat{I}(\boldsymbol{p}_t) = \{\sigma_{\text{SLS}}(\boldsymbol{p}_i, \psi_i)\cos(\theta_i)\rho_i g\}^{1/\gamma} \tag{9}$$

where $\boldsymbol{p}_t$ is the pixel coordinate corresponding to $\boldsymbol{p}_i$; g is the automatic gain control; ρ_i is the reflectance at the position of $\boldsymbol{p}_i$; γ is the gamma correction of the camera. Therefore, according to the above equation, the rendered pixel value at corresponding positions between the target view and source view has the following relationship:

$$\begin{aligned} k_{s\to t}(\boldsymbol{p}_t, \boldsymbol{p}_s) &= \frac{\hat{I}_t^{\gamma}(\boldsymbol{p}_t)}{\hat{I}_s^{\gamma}(\boldsymbol{p}_s)} \\ &= \frac{\sigma_{SLS}(\boldsymbol{p}_i, \psi_i)\cos(\theta_i)}{\sigma_{SLS}(\boldsymbol{p}_j, \psi_j)\cos(\theta_j)} \cdot \frac{\rho_i}{\rho_j} \cdot \frac{g_t}{g_s} \end{aligned} \tag{10}$$

where $\hat{I}_t$ and $\hat{I}_s$ represent the color captured by the camera of the source frame and the target frame; $k_{s\to t}$ is the photometric ratio of the target image to the source image; $\boldsymbol{p}_t$ and $\boldsymbol{p}_s$ are the corresponding points in the target frame and the source frame, while $\boldsymbol{p}_j$ is the spatial 3D point corresponding to $\boldsymbol{p}_s$. To simplify the expression, let:

$$\begin{aligned} R_t(\boldsymbol{p}_i) = \sigma_{SLS}(\boldsymbol{p}_i, \psi_i)\cos(\theta_i) \\ R_s(\boldsymbol{p}_j) = \sigma_{SLS}(\boldsymbol{p}_j, \psi_j)\cos(\theta_j) \end{aligned} \tag{11}$$

where R_t and R_s represent the radiance received by the target frame and the source frame, respectively. Additionally, since the reflectance at the same position is the same: $\rho_i = \rho_j$, Equation 11 can be simplified as:

$$\begin{aligned} k_{s\to t}(\boldsymbol{p}_t, \boldsymbol{p}_s) &= \frac{R_t(\boldsymbol{p}_i)}{R_s(\boldsymbol{p}_j)} \cdot \frac{g_t}{g_s} \\ &= k_R(\boldsymbol{p}_t, \boldsymbol{p}_s) \cdot k_g \end{aligned} \tag{12}$$

where k_R is the radiance ratio; k_g is the automatic gain control ratio. Therefore, there is the following relationship between the target rendered image and the source rendered image: $\hat{I}_t = k_{s\to t}^{1/\gamma}\hat{I}_s$. Rendered images are simulations of images captured by cameras, thus the original images captured by the cameras, I_t, I_s also follow this relationship:

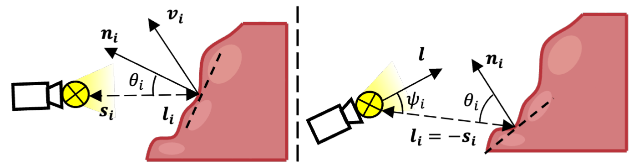

Figure 2: (a) and (b) are the specular reflection model and spotlight illumination model in endoscopy. Here, $\boldsymbol{n}_i$ represents the normal vector of the object's surface, derived from the depth map; $\boldsymbol{s}_i$ is the unit vector in the viewing direction; $\boldsymbol{l}_i$ represents the unit vector in the direction of light incidence; $\boldsymbol{v}_i$ is the unit vector in the direction of light reflection.

$$I_t = k_{s\to t}^{1/\gamma} I_s = (k_R \cdot k_g)^{1/\gamma} I_s \tag{13}$$

Therefore, the key to photometric alignment between the target view and the source view lies in calculating k_R and k_g. Here, k_R represents the radiance ratio captured by the camera, that is, the ratio between R_t and R_s. k_g is the ratio of automatic exposure gain between the target view and the source view. Combining with Equation 9, write R_t as:

$$\begin{aligned} R_t(\boldsymbol{p}_i) &= \frac{\sigma_0}{||\boldsymbol{p}_i||^2} e^{-\mu(1-cos(\psi))} cos(\theta_i) \\ &= \frac{\sigma_0}{||\boldsymbol{p}_i||^2} e^{-\mu(1-z(\overline{\boldsymbol{p}}_i))} (-(\overline{\boldsymbol{p}}_i \cdot \boldsymbol{n}_i)) \end{aligned} \tag{14}$$

where the normal vector $\boldsymbol{n}_i$ can be derived from the depth map; $\overline{\boldsymbol{p}}_i$ is the unit vector corresponding to $\boldsymbol{p}_i$. Similarly, due to the existence of the normal vectors of the source frame and the target frame being $\boldsymbol{n}_s = \mathbf{T}_{t\to s}\boldsymbol{n}_t$, where $\mathbf{T}_{t\to s}$ is camera pose from target frame to source frame, the calculation of R_s is similar to the above. Finally, the radiance ratio is calculated.

To compute automatic gain control ratio k_g, we rewrite Equation 14:

$$I_t^{\gamma} = k_g(k_R \cdot I_s^{\gamma}) \tag{15}$$

Therefore, after removing the gamma gain, a proportional relationship exists between the corresponding pixels of the target view and the source view, with k_g as the proportional coefficient. Consequently, k_g is calculated as follows:

$$k_g = \frac{\sum_{\boldsymbol{p}_t} I_t^{\gamma}(\boldsymbol{p}_t)}{\sum_{\boldsymbol{p}_s} k_R(\boldsymbol{p}_t, \boldsymbol{p}_s) I_s^{\gamma}(\boldsymbol{p}_s)} \tag{16}$$

where $\boldsymbol{p}_t$ and $\boldsymbol{p}_s$ are the pixel coordinates corresponding to the target view and source view, respectively. With the above, $k_{s\to t}$ can be calculated for photometric alignment.

Aligned Wrap. The purpose of this component is to generate a synthetic frame with the same photometric values as the target frame. First, we use the depth map and ego-motion to wrap the source frame into a synthetic frame (Jaderberg et al. 2015; Bian et al. 2019). Then, we apply the photometric ratio $k_{s\to t}^{1/\gamma}$ between the target frame and the source frame to the synthetic frame, as shown in the following equation:

$$\tilde{I}_t^s = k_{s\to t}^{1/\gamma} I_t^s \tag{17}$$

Method	Dataset	Error ↓				Accuracy ↑	
		AbsRel	SqRel	RMSE	RMSE log	< 1.25	$< 1.25^2$
Monodepth2 (Godard et al. 2019)	C	0.297	1.642	18.64	0.392	0.489	0.731
SC-Depth (Bian et al. 2019)	C	0.084	0.463	4.062	0.123	0.937	0.989
Endo-SfM (Ozyoruk et al. 2021)	C	0.249	10.583	7.065	0.219	0.815	0.933
AF-SfM (Shao et al. 2022)	C	0.202	4.045	6.768	0.238	0.756	0.929
LightDepth (Rodríguez-Puigvert et al. 2023)	C	0.081	1.810	6.550	0.227	0.928	0.981
MonoLoT (He et al. 2024)	C	0.096	1.967	4.321	0.178	0.934	0.895
DepthAnything (Yang et al. 2024a)	C	0.219	2.741	8.648	0.253	0.665	0.899
Monodepth2*	C	0.082	0.379	3.536	0.116	0.947	0.991
PC-Depth	C	**0.076**	**0.333**	**3.279**	**0.110**	**0.949**	**0.991**
SfMLearner (Zhou et al. 2017)	S	0.086	1.021	7.553	0.121	0.925	0.987
Monodepth2 (Godard et al. 2019)	S	0.066	0.577	5.781	0.093	0.961	0.995
SC-Depth (Bian et al. 2019)	S	0.064	0.651	6.075	0.096	0.957	0.992
Endo-SfM (Ozyoruk et al. 2021)	S	0.070	0.725	6.410	0.099	0.956	0.992
AF-SfM (Shao et al. 2022)	S	0.065	0.557	5.459	0.089	0.958	**0.997**
MVDE (Li et al. 2023)	S	0.066	0.655	6.441	0.086	0.955	0.993
Yang et al. (Yang et al. 2024b)	S	0.062	0.558	5.985	0.090	0.962	0.993
DepthAnything (Yang et al. 2024a)	S	0.068	0.590	5.643	0.092	0.964	0.994
Monodepth2*	S	0.063	0.568	5.695	0.091	0.961	0.995
PC-Depth	S	**0.057**	**0.514**	**5.362**	**0.084**	**0.971**	0.994

Table 1: Quantitative depth comparison on the C3VD and SCARED datasets. ↑ denotes higher the better; ↓ denotes lower the better; The best performance in each block is highlighted as bold; C denotes C3VD dataset; S denotes SCARED dataset; * denotes results based on our photometric alignment.

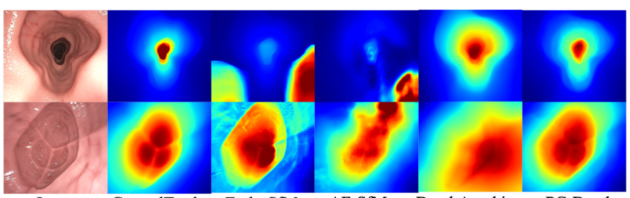

Figure 3: Qualitative depth comparison on the C3VD dataset.

where $\tilde{I}_t^s$ is the synthetic frame after photometric alignment. Finally, the discrepancy between target frame and synthetic frame is used to compute the Photometric Loss.

Photometric Loss

Photometric loss is employed to measure the difference between the synthetic frame after photometric alignment and the target frame (Yin and Shi 2018; Ranjan et al. 2019). Meanwhile, following (Bian et al. 2019), we add an additional image dissimilarity loss SSIM (Wang et al. 2004) for better handling complex illumination changes. We formulate the function as:

$$\mathcal{L}_P = \frac{1}{|V|}\sum_{\boldsymbol{p}\in V}(\lambda_i\|I_t(\boldsymbol{p}) - \tilde{I}_t^s(\boldsymbol{p})\|_1 + \lambda_s(\frac{1-\mathrm{SSIM}_{ta}(\boldsymbol{p})}{2})) \tag{18}$$

where V stands for valid points that are successfully projected from I_s to the image plane of I_t, and $|V|$ defines the number of points in V; SSIM_{ta} stands for the element-wise similarity between I_t and $\tilde{I}_t^s$ by the SSIM function. Following (Bian et al. 2019; Yin and Shi 2018), we choose L_1 loss and use $\lambda_i = 0.15$ and $\lambda_s = 0.85$ in our framework. Since highlight areas do not satisfy the Lambertian assumption, and Photometric Match module is based on this assumption, highlight areas are excluded when calculating photometric loss. We modify the Photometric Loss term as:

$$\begin{aligned} \mathcal{L}_P^{MH} &= \frac{1}{|M|}\sum_{p\in M}(M(p)\cdot\mathcal{L}_P(p)) \\ M &= M_{ht}\cap M_{hs}\cap V \end{aligned} \tag{19}$$

where M_{ht} and M_{hs} represent the pixels that are not highlights on the target and source frame, which detected by (Arnold et al. 2010).

Smoothness Loss and Geometry Consistency Loss

Smoothness Loss. To address the issue of gradient loss in missing texture regions, we introduced an edge-aware Smoothness Loss $\mathcal{L}_S$ (Ranjan et al. 2019). However, significant gradients at the edges of highlight areas can distort the Smoothness Loss calculation. To mitigate the impact of these highlights, we employ the approach (Arnold et al. 2010) to remove highlights from the image before computing the Smoothness Loss. This process is formulated as follows:

$$\mathcal{L}_S = \sum_{\boldsymbol{p}}\left(e^{\nabla I_t^H(\boldsymbol{p})}\cdot\nabla D_t(\boldsymbol{p})\right)^2 \tag{20}$$

Method	AbsRel	SqRel	RMSE	RMSE log
SfMLearner	0.151	3.917	17.451	0.191
Monodepth2	0.123	2.205	12.927	0.152
SC-Depth	0.116	2.015	12.415	0.149
Endo-SfM	0.117	2.120	12.970	0.151
AF-SfM	0.140	3.151	15.371	0.174
DepthAnything	0.139	4.117	18.282	0.172
PC-Depth	0.104	1.663	11.394	0.131

Table 2: Quantitative depth comparison on the SERV-CT dataset. All methods are self-supervised monocular trained on the SCARED dataset.

where ∇ represents the first derivative along spatial directions, ensuring that smoothness is guided by the edges of the images. $I_t^H(\boldsymbol{p})$ denotes the target frame with highlights removed.

Geometry Consistency Loss. Geometric Consistency Loss constrains the similarity between depth maps, ensuring they represent a three-dimensional structure with the same scale (Bian et al. 2019). Specifically, according to the camera imaging model, the depth map of the source frame can be expressed as:

$$D_s(\boldsymbol{p}_s) = (T_{t\to s} D_t(\boldsymbol{p}_t)\mathbf{K}^{-1}\boldsymbol{p}_t)_z \tag{21}$$

where $\boldsymbol{p}_t$ and $\boldsymbol{p}_s$ are the pixel coordinates of the target and source frames; z indicates that only the z-axis coordinates are considered; $\mathbf{K}$ represents the camera intrinsic matrix. The Geometry Consistency Loss is formulated as follows:

$$\begin{aligned} D_{diff} &= \frac{|D_s - \hat{D}_s|}{D_s + \hat{D}_s} \\ \mathcal{L}_{GC} &= \frac{1}{|V|}\sum_{\boldsymbol{p}\in V} D_{diff}(\boldsymbol{p}) \end{aligned} \tag{22}$$

where $\hat{D}_s$ is the interpolation map of D_s. By minimizing the depth inconsistency among a batch of samples, consistency can naturally propagate throughout the entire sequence.

Experiments

Datasets and metrics

C3VD (Bobrow et al. 2023). The images are captured by using a genuine Olympus CF-HQ190L endoscope within a silicone model of a human colon. Since the silicone material is opaque, the only light source available is in the endoscope. The C3VD dataset consists of 22 video sequences, totaling 10,015 frames. We allocate 8,690 frames for training, 148 for validation, and 2,888 are designated for testing purposes.

SCARED (Allan et al. 2021). The dataset is acquired using a da Vinci Xi endoscope on fresh porcine cadaver abdominal anatomy, comprising 35 endoscopic videos totaling 22,950 frames. Following (Shao et al. 2022), the dataset is divided into a training set containing 15,351 frames, a validation set containing 1,705 frames, and a test set containing 551 frames.

Method	$\text{ATE}_{Trans} \downarrow$	$\text{ATE}_{Rot} \times 1e^{-2} \downarrow$
SC-Depth	1.6472	3.0331
Endo-SfM	2.6360	2.9322
AF-SfM	0.3231	**2.1728**
PC-Depth	**0.2660**	2.4513

Table 3: Quantitative average performance of ego-motion on trans-t2-a from C3VD dataset.

SERV-CT (Edwards et al. 2022). SERV-CT includes 16 stereo pairs collected from ex vivo porcine torso cadavers, along with corresponding depth and disparity ground truth.

Evaluation metrics. Similar to previous work (Bian et al. 2019; Yang et al. 2024a), we adhere to the standard evaluation protocol to validate the effectiveness of our proposed method in our experiments. This involves assessing the relative absolute error (Abs Rel), relative squared error (Sq Rel), root mean squared error (RMSE), root mean squared logarithmic error (RMSE log), threshold accuracy (with thresholds of $\delta < 1.25$ and $\delta < 1.25^2$), Absolute Trajectory Translation Error (ATE_{Trans}) and Absolute Trajectory Rotation Error (ATE_{Rot}).

Implementation details

Network architecture. The depth network employs an encoder-decoder structure with skip connections, utilizing ResNet-18 (He et al. 2016) as the encoder. The design of the decoder follows (Zhou et al. 2017; Bian et al. 2019). For the pose network, we adopt the model from SC-Depth (Bian et al. 2019). The depth network takes a single RGB image as input and outputs a depth map, while the pose network estimates a 6D relative camera pose from a concatenated pair of RGB images. To simplify computations and enhance accuracy, we avoid multi-scale loss calculations (Zhou et al. 2017; Godard et al. 2019), and instead achieve superior results using a single scale.

Training setup. Our PC-Depth is trained on one GTX 3090 GPU with a batch size of 4 for 20 epochs. Following (Bian et al. 2019), adam optimizer (Kingma and Ba 2014) is used with an initial learning rate of 1e-4 and drops to 1e-5 after 10 epochs. We utilize the pre-trained weights on ImageNet (Deng et al. 2009) for ResNet initialization. We adopt α = 1.0, β = 0.01, γ = 0.1 and ω = 0.1 in Equation 1. For fairness, all comparison methods use the same settings.

Comparisons with the state-of-the-art

Depth results. As shown in Table 1, our PC-Depth outperforms previous methods on the C3VD and SCARED datasets. Specifically, compared with the recent DepthAnything, which leverages large-scale data coverage for robust monocular depth estimation, PC-Depth demonstrates significant improvements of 0.143%, 2.408%, 5.369%, and 0.143% in AbsRel, SqRel, RMSE, and RMSE log on the C3VD benchmark, respectively. The severe photometric fluctuations, induced by local and global photometric inconsistencies, tend to create a highly biased supervisory signal.

$\mathcal{L}_H$	$\mathcal{L}_P$	$\mathcal{L}_S$	Error ↓ AbsRel	SqRel	RMSE	RMSE log	Accuracy ↑ < 1.25	$< 1.25^2$
✓	×	×	0.091	0.475	4.148	0.128	0.931	0.988
×	✓	×	0.079	0.348	3.368	0.114	0.949	0.989
✓	×	✓	0.091	0.457	3.950	0.125	0.930	0.990
✓	✓	×	0.082	0.351	3.431	0.116	0.947	0.989
✓	✓	✓	**0.076**	**0.333**	**3.279**	**0.110**	**0.949**	**0.991**

Table 4: Ablation study on the photometric consistent of depth setimation on C3VD dataset. $\mathcal{L}_H$ is Highlight Loss; $\mathcal{L}_P$ is Photometric Loss with Photometric Match; $\mathcal{L}_S$ is Smoothness Loss with Highlight Removal.

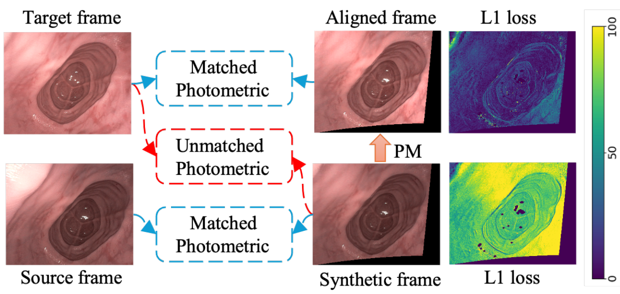

Figure 4: Ablation study of the Photometric Match (PM) module. PM effectively mitigates ambiguous supervision resulting from global photometric inconsistency, thereby reducing the L1 loss.

We believe that a photometric consistency constraint can offer an advantage in such scenarios. It is worth noting that Endo-SfM introduces an affine transformer to tackle the issue of photometric inconsistency. However, this approach faces challenges in adapting to complex lighting environments, resulting in sub-optimal performance.

Generalization Robustness. We directly validate the models trained by the SCARED on the SERV-CT dataset as shown in Table 2. Our PC-Depth achieves superior results than the other methods, revealing its strong generalization ability. Additionally, we use Monodepth2 (Godard et al. 2019) as baseline, a relatively basic depth model, and achieve better results. We believe the results can be further enhanced if equipped with more advanced architectures.

Qualitative results. We present some depth predictions from our PC-Depth and other methods on the C3VD datasets. As shown in Figure 3, our model provides more accurate depth estimations compared to other methods. In contrast, Endo-SfM shows significant errors in depth estimation in regions with photometric variations.

Ego-motion results. Table 3 presents a quantitative comparison on the C3VD dataset, demonstrating that our method achieves a lower ATE_{Trans} than several typical unsupervised methods. However, the accuracy improvement in ego-motion estimation with our method is not as pronounced as in depth estimation. This difference can be attributed to the nature of the tasks: depth estimation requires pixel-to-pixel mapping, where local and global photometric inconsistencies can significantly impact the accuracy of the depth map.

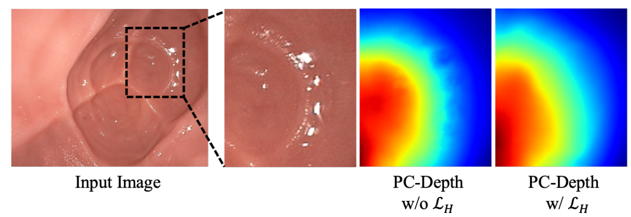

Figure 5: Ablation study of the Highlight Loss ($\mathcal{L}_H$) on the C3VD datasets.

Ablation Studies

As shown in Table 4, we ablate the effectiveness of both global and local photometric consistencies on depth estimation. Specifically, we observe that the Highlight Loss effectively suppresses local photometric inconsistencies, leading to improved model performance. Meanwhile, we find that the model incorporates Photometric Match, the computation of Photometric Loss becomes more reliable. As illustrated in Figure 4, the L1 loss of the aligned synthetic frame is significantly smaller than that of the unaligned synthetic frame. This improvement can be attributed to the Photometric Match (PM) module, which adjusts the photometric values of the synthetic frame to match the target frame. Furthermore, as depicted in Figure 5, incorporating Highlight Loss into the network results in smoother depth predictions in highlight regions. This enhancement arises from the Highlight Loss constraining the surface normal vectors within the highlight regions, aligning the reflected light with the observation direction, and thereby providing reliable supervisory signals for these areas.

Conclusion

In this paper, we propose PC-Depth, a simple yet effective method for maintaining photometric-consistent depth estimation in endoscopy. The core of our approach leverages the unique properties of endoscopic imaging to obtain a robust supervision signal. Additionally, we introduce a photometric alignment strategy that can be easily adapted to other baselines, ensuring photometric consistency. Comprehensive experiments demonstrate that PC-Depth significantly outperforms previous state-of-the-art methods on three datasets.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFA0714003, in part by the Science and Technology Planning Project of Sichuan Province under Grant 2021YFQ0059, and in part by the National Natural Science Foundation of China under Grant No. 61471250.

References

Allan, M.; Mcleod, J.; Wang, C.; Rosenthal, J. C.; Hu, Z.; Gard, N.; Eisert, P.; Fu, K. X.; Zeffiro, T.; Xia, W.; et al. 2021. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*.

Arnold, M.; Ghosh, A.; Ameling, S.; and Lacey, G. 2010. Automatic segmentation and inpainting of specular highlights for endoscopic imaging. *EURASIP Journal on Image and Video Processing*, 2010: 1–12.

Baker, S.; and Matthews, I. 2004. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56: 221–255.

Bian, J.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.-M.; and Reid, I. 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *NeurIPS*, 32.

Bobrow, T. L.; Golhar, M.; Vijayan, R.; Akshintala, V. S.; Garcia, J. R.; and Durr, N. J. 2023. Colonoscopy 3D video dataset with paired depth from 2D-3D registration. *Medical image analysis*, 90: 102956.

Cao, Y.; Wu, Z.; and Shen, C. 2017. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11): 3174–3182.

Chen, L.; Tang, W.; John, N. W.; Wan, T. R.; and Zhang, J. J. 2018. SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. *Computer methods and programs in biomedicine*, 158: 135–146.

Chen, Y.; Schmid, C.; and Sminchisescu, C. 2019. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, 7063–7072.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Edwards, P. E.; Psychogyios, D.; Speidel, S.; Maier-Hein, L.; and Stoyanov, D. 2022. SERV-CT: A disparity dataset from cone-beam CT for validation of endoscopic 3D reconstruction. *Medical image analysis*, 76: 102302.

Edwards, P. E.; Chand, M.; Birlo, M.; and Stoyanov, D. 2021. The challenge of augmented reality in surgery. *Digital Surgery*, 121–135.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27.

Fitzpatrick, J. M. 2010. The role of registration in accurate surgical guidance. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 224(5): 607–622.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *ICCV*, 3828–3838.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, L.; Wang, G.; and Hu, Z. 2018. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 27(9): 4676–4689.

He, Q.; Feng, G.; Bano, S.; Stoyanov, D.; and Zuo, S. 2024. MonoLoT: Self-Supervised Monocular Depth Estimation in Low-Texture Scenes for Automatic Robotic Endoscopy. *IEEE Journal of Biomedical and Health Informatics*.

Horn, B. K.; and Schunck, B. G. 1981. Determining optical flow. *Artificial intelligence*, 17(1-3): 185–203.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *NeurIPS*, 28.

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

Li, L.; Li, X.; Yang, S.; Ding, S.; Jolfaei, A.; and Zheng, X. 2020. Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery. *IEEE Transactions on Industrial Informatics*, 17(6): 3920–3928.

Li, S.; Tu, Y.; Gong, Y.; Zhong, B.; and Li, Z. 2024a. Dual Attention Encoder with Joint Preservation for Medical Image Segmentation. In *ECAI*, 330–337.

Li, S.; Tu, Y.; Xiang, Q.; and Li, Z. 2024b. MAGIC: Rethinking Dynamic Convolution Design for Medical Image Segmentation. In *ACM MM*, 9106–9115.

Li, W.; Hayashi, Y.; Oda, M.; Kitasaka, T.; Misawa, K.; and Mori, K. 2023. Multi-view Guidance for Self-supervised Monocular Depth Estimation on Laparoscopic Images via Spatio-Temporal Correspondence. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 429–439. Springer.

Liu, F.; Shen, C.; Lin, G.; and Reid, I. 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10): 2024–2039.

Liu, X.; Sinha, A.; Ishii, M.; Hager, G. D.; Reiter, A.; Taylor, R. H.; and Unberath, M. 2019. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE transactions on medical imaging*, 39(5): 1438–1447.

Mahjourian, R.; Wicke, M.; and Angelova, A. 2018a. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 5667–5675.

Mahjourian, R.; Wicke, M.; and Angelova, A. 2018b. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 5667–5675.

Modrzejewski, R.; Collins, T.; Hostettler, A.; Marescaux, J.; and Bartoli, A. 2020. Light modelling and calibration in laparoscopy. *International journal of computer assisted radiology and surgery*, 15: 859–866.

Ozyoruk, K. B.; Gokceler, G. I.; Bobrow, T. L.; Coskun, G.; Incetan, K.; Almalioglu, Y.; Mahmood, F.; Curto, E.; Perdigoto, L.; Oliveira, M.; et al. 2021. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical image analysis*, 71: 102058.

Ranjan, A.; Jampani, V.; Balles, L.; Kim, K.; Sun, D.; Wulff, J.; and Black, M. J. 2019. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 12240–12249.

Recasens, D.; Lamarca, J.; Fácil, J. M.; Montiel, J.; and Civera, J. 2021. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4): 7225–7232.

Ren, Z.; He, T.; Peng, L.; Liu, S.; Zhu, S.; and Zeng, B. 2017. Shape recovery of endoscopic videos by shape from shading using mesh regularization. In *Image and Graphics: 9th International Conference, ICIG 2017, Shanghai, China, September 13-15, 2017, Revised Selected Papers, Part III 9*, 204–213. Springer.

Rodríguez-Puigvert, J.; Batlle, V. M.; Montiel, J.; Martinez-Cantin, R.; Fua, P.; Tardós, J. D.; and Civera, J. 2023. LightDepth: Single-View Depth Self-Supervision from Illumination Decline. In *ICCV*, 21273–21283.

Shao, S.; Pei, Z.; Chen, W.; Zhu, W.; Wu, X.; Sun, D.; and Zhang, B. 2022. Self-supervised monocular depth and egomotion estimation in endoscopy: Appearance flow to the rescue. *Medical image analysis*, 77: 102338.

Taylor, R. H.; Menciassi, A.; Fichtinger, G.; Fiorini, P.; and Dario, P. 2016. Medical robotics and computer-integrated surgery. *Springer handbook of robotics*, 1657–1684.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; and Huang, Q. 2024a. SMART: Syntax-Calibrated Multi-Aspect Relation Transformer for Change Captioning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(07): 4926–4943.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2024b. Context-aware Difference Distilling for Multichange Captioning. In *ACL*, 7941–7956.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Zhou, Y.; He, S.; Li, T.; Huang, F.; Ding, Q.; Feng, X.; Liu, M.; and Li, Q. 2024. MonoPCC: Photometric-invariant Cycle Constraint for Monocular Depth Estimation of Endoscopic Images. *arXiv preprint arXiv:2404.16571*.

Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; and Sebe, N. 2017. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, 5354–5362.

Xu, D.; Wang, W.; Tang, H.; Liu, H.; Sebe, N.; and Ricci, E. 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In *CVPR*, 3917–3925.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, J.; and Yang, J. 2023. Desnet: Decomposed scale-consistent network for unsupervised depth completion. In *AAAI*, volume 37, 3109–3117.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *CVPR*, 10371–10381.

Yang, Z.; Pan, J.; Dai, J.; Sun, Z.; and Xiao, Y. 2024b. Self-Supervised Lightweight Depth Estimation in Endoscopy Combining CNN and Transformer. *IEEE Transactions on Medical Imaging*, 43(5): 1934–1944.

Yang, Z.; Wang, P.; Wang, Y.; Xu, W.; and Nevatia, R. 2018. Lego: Learning edge with geometry all at once by watching videos. In *CVPR*, 225–234.

Yin, W.; Liu, Y.; Shen, C.; and Yan, Y. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 5684–5693.

Yin, Z.; and Shi, J. 2018. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 1983–1992.

Zhao, C.; Zhang, Y.; Poggi, M.; Tosi, F.; Guo, X.; Zhu, Z.; Huang, G.; Tang, Y.; and Mattoccia, S. 2022. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *3DV*, 668–678.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 2881–2890.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 1851–1858.

Zhuang, C.; Lu, Z.; Wang, Y.; Xiao, J.; and Wang, Y. 2022. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *AAAI*, volume 36, 3653–3661.

Zou, Y.; Luo, Z.; and Huang, J.-B. 2018. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 36–53.