

DeLightMono: Enhancing Self-Supervised Monocular Depth Estimation in Endoscopy by Decoupling Uneven Illumination

Mingyang Ou¹, Haojin Li¹, Yifeng Zhang¹, Ke Niu², Zhongxi Qiu³, Heng Li⁴, Jiang Liu^{3,1}

¹Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

²Computer School, Beijing Information Science and Technology University, Beijing 100192, China

³Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen, China

⁴Faculty of Biomedical Engineering, Shenzhen University of Advanced Technology, Shenzhen, China

liheng@suat-sz.edu.cn

Abstract

Self-supervised monocular depth estimation serves as a key task in the development of endoscopic navigation systems. However, performance degradation persists due to uneven illumination inherent in endoscopic images, particularly in low-intensity regions. Existing low-light enhancement techniques fail to effectively guide the depth network. Furthermore, solutions from other fields, like autonomous driving, require well-lit images, making them unsuitable and increasing data collection burdens. To this end, we present DeLightMono - a novel self-supervised monocular depth estimation framework with illumination decoupling. Specifically, endoscopic images are represented by a designed illumination-reflectance-depth model, and are decomposed with auxiliary networks. Moreover, a self-supervised joint-optimizing framework with novel losses leveraging the decoupled components is proposed to mitigate the effects of uneven illumination on depth estimation. The effectiveness of the proposed methods was rigorously verified through extensive comparisons and an ablation study performed on two public datasets.

Code — <https://github.com/ComgLq24/AAAI-2026-DeLightMono>

1 Introduction

Monocular endoscopy constitutes a fundamental component of disease diagnosis and intervention in minimally invasive surgery (Zhu et al. 2021). Nevertheless, the confined anatomical structure of the human body greatly restricts the scope of view in monocular endoscopy and further increases the difficulty of intra-operative manipulation (Yue et al. 2023). To alleviate the situation, surgical navigation (Metzger et al. 2024) and robotic-assisted surgery (Lu et al. 2023) systems have been extensively researched to provide more perception and control under endoscopy. While these systems ease the burden of surgeons, they both heavily rely on precisely estimating the depth.

Depth estimation from a sequence of frames has been a long-standing task in computer vision. Traditional methods such as structure from motion (SfM) (Enqvist, Kahl, and Olsson 2011) and simultaneous localization and mapping (SLAM) (Engel, Schöps, and Cremers 2014) have been

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

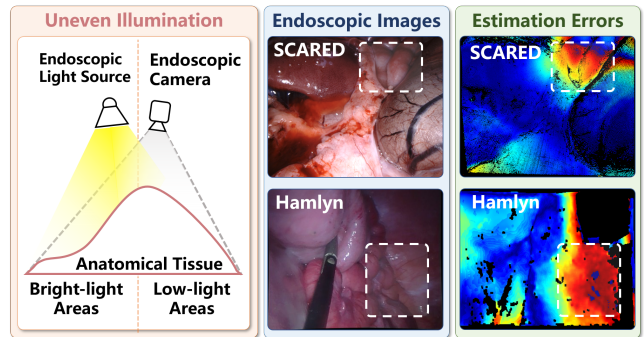


Figure 1: Uneven illumination affects depth estimation in endoscopic images. Left: illustration of uneven illumination. Middle: examples from SCARED and Hamlyn datasets. Right: error maps of Monodepth2, black areas contain nan depth label, where the highlighted regions suffer from elevated depth estimation errors.

widely adopted for this purpose. Recently, deep learning-based methods have shown remarkable potential in addressing the depth estimation task (Fu et al. 2018; Karsch, Liu, and Kang 2014). Pioneering works (Li et al. 2015; Eigen, Puhrsch, and Fergus 2014) train the depth net in a supervised manner. However, in particular domains such as endoscopy, labeled data acquisition comes at a high cost, as the scene needs to be stationary while obtaining the ground truth.

Accordingly, self-supervised monocular depth estimation (Zhou et al. 2017) is proposed to mitigate the need for labeled data. The depth network is trained by aligning 3D correspondences between the target frame and temporal adjacent source frames using depth cues. Building upon this, a series of improvements (Godard et al. 2019; Lyu et al. 2021; Amitai, Klein, and Treibitz 2023) have been made by mining more informative geometric priors, scene-specific features to enhance depth estimation accuracy and generalization. As for endoscopic images, researchers seek solutions to deal with challenges such as complex camera motion (Li et al. 2022), specular reflection (Li et al. 2024), textureless anatomy (He et al. 2024), etc.

However, these methods still struggle with precise depth estimation in endoscopic scenarios. As illustrated in Fig. 1,

monocular depth estimation suffers from uneven illumination in endoscopic images. Particularly, the estimation error increases in the low-light regions. Furthermore, the illumination level also varies across datasets, leading to a more challenging generalization problem.

While low-light enhancement methods (Ma et al. 2022; Wu et al. 2023) improve image visibility, offering a viable solution, they do not guide the depth net to learn the illumination distribution, resulting in suboptimal improvement. Moreover, there exist methods (Vankadari et al. 2023; Zheng et al. 2023) facing a similar illumination problem in the natural scene; they require external resources such as bright light images or their corresponding depth maps during training, which is infeasible for endoscopy.

For these concerns, we propose a novel self-supervised framework (DeLightMono) with uneven illumination decoupling to enhance monocular depth estimation in endoscopic images. Firstly, we observe that endoscopic images are predominantly illuminated by a single artificial light source. The insight drives our Illumination-Reflectance-Depth (IRD) model, which decouples the illumination-invariant reflectance network from the image. We then introduce a unified self-supervised framework for jointly optimizing the IRD decomposition and monocular depth estimation. Specifically, the IRD decoupling branch involves a reconstruction loss and an intensity ratio consistency loss with two auxiliary networks to correctly estimate illumination and reflectance. On the other hand, the depth estimation branch benefits from a redesigned photometric loss and an illumination-map guided degradation consistency loss, both aimed at mitigating uneven illumination. Our major contributions are as follows:

- To alleviate the negative impact of uneven illumination on monocular depth estimation, we design an illumination-reflectance-depth (IRD) modeling for endoscopic images.
- Building on this modeling, a jointly optimized self-supervised framework is proposed for depth estimation and IRD decoupling, where the decoupled IRD components guide the depth net to correctly estimate the depth from uneven illumination by four novel losses.
- The robustness of the proposed framework against uneven illumination is verified by a comprehensive comparison with state-of-the-art methods and an ablation study on two public endoscopic depth datasets.

2 Related Work

2.1 Self-Supervised Monocular Depth Estimation

Predicting 3D depth information from 2D images has been a fascinating topic in 3D vision. Early-stage methods (Eigen, Puhrsch, and Fergus 2014; Li et al. 2015; Karsch, Liu, and Kang 2014) leveraged annotated datasets to supervise the training of the depth estimation model.

As obtaining labeled depth data is expensive, Godard, Mac Aodha, and Brostow (2017) took the first step toward alleviating this issue by enforcing the left-right consistency across unlabeled stereo images. Zhou et al. (2017) extended

the idea to monocular settings by employing an auxiliary camera pose estimation network. Subsequent works focus on fusing higher-resolution features (Lyu et al. 2021; Zhao et al. 2022) and developing lightweight architectures for edge devices (Zhao et al. 2022). In addition, various cues such as dynamic and static objects (Klingner et al. 2020; Godard et al. 2019; Zhou et al. 2025) and depth uncertainty (Poggi et al. 2020) were incorporated in the training process to improve overall performance.

Unlike natural images, an endoscopic scene poses greater challenges for depth estimation. Some methods introduced keypoint matching losses (He et al. 2024) or fused global-local features (Fan et al. 2024) to address the distortion of low-texture regions. Others employed the Siamese network (Li et al. 2022) and transformation consistency module (Yue and Gu 2023) to learn a more accurate camera pose. As varying specular reflection breaks the photometric consistency assumption, researchers adopted appearance flow (Shao et al. 2022; Ozyoruk et al. 2021) networks and retinex theory (Li et al. 2024) to mitigate the effects. In recent years, several studies have also explored transferring depth estimation foundation models (Yang et al. 2024; Ranftl et al. 2020) to endoscopic datasets (Cui et al. 2024; Budd and Vercauteren 2024).

However, existing methods rarely consider the negative impact of spatially uneven illumination in each frame (as in Fig. 1), which prohibits the depth net from learning the correct depth in the low-light regions. Moreover, the appearance of low-light regions remains stable across adjacent frames, yet depth accuracy drops significantly in these regions. Consequently, prevalent methods that study varying specular reflection are insufficient to address this issue.

2.2 Low-Light Image Enhancement

The objective of low-light image enhancement is to improve the image quality by recovering visual details in the dark regions. Traditional methods like CLAHE (Pizer et al. 1987; Reza 2004) and Lime (Guo, Li, and Ling 2016) calculated the statistical information or optimized an illumination map with a mathematical model to redistribute the pixel intensity. However, this type of method can amplify noise in images, especially in uniform or low-contrast regions, leading to a degradation in overall image quality. Additionally, their effectiveness is highly sensitive to parameters. Alternative data-driven approaches incorporated the Retinex theory to decouple illumination and reflectance maps, thereby enabling effective illumination correction (Zhang et al. 2021; Chen et al. 2018; Cai et al. 2023). These methods all require a paired normal-/low-light images dataset for supervised training, whereas most recently, Ma et al. (2022) and Guo et al. (2020) developed unsupervised methods and iteratively lit-up the low-light images.

3 Method

This section first reviews the conventional approach to self-supervised monocular depth estimation (Sec. 3.1). Noting its poor performance in endoscopy due to uneven illumination, we introduce Illumination-Reflectance-Depth (IRD) modeling to explicitly connect illumination, reflectance, and depth.

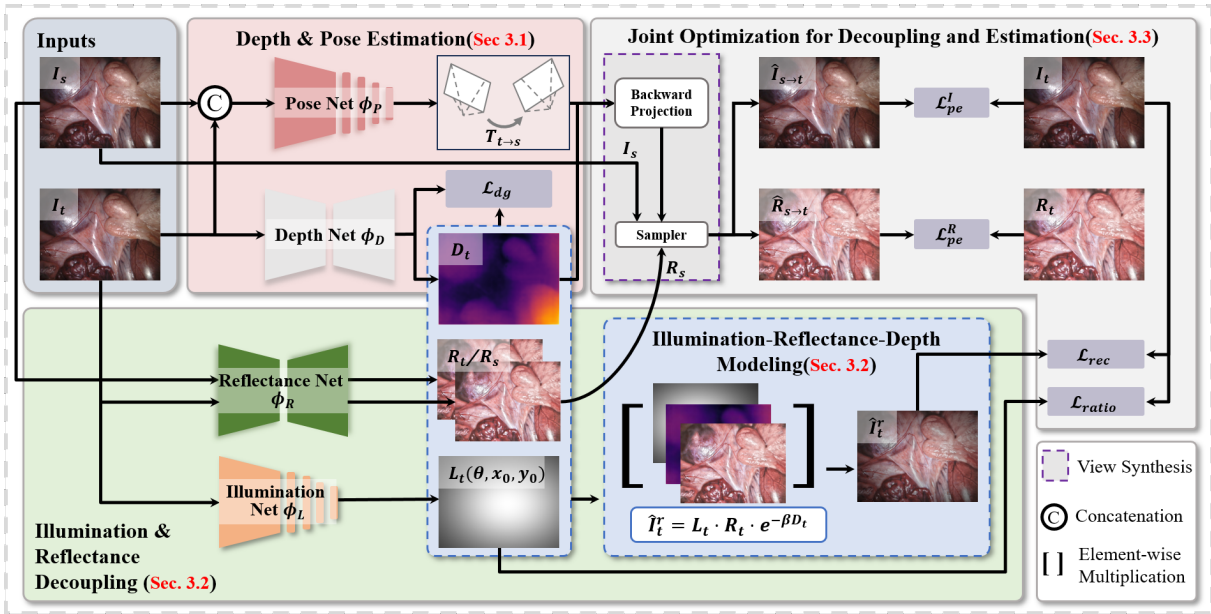


Figure 2: Overview of the proposed method.

Based on this model, we design an auxiliary branch with two networks to decouple the illumination and reflectance maps from an endoscopic image (Sec. 3.2). Finally, we propose a joint self-supervised framework that employs novel loss functions to mutually optimize the IRD decoupling branch and the depth estimation branch (Sec. 3.3).

3.1 Self-Supervised Monocular Depth Estimation

Following (Godard et al. 2019), self-supervised monocular depth estimation learns the depth from the reconstruction loss between the target frame and a synthesized view from the adjacent frames. As shown in the red block of Fig. 2, given a target endoscopic frame I_t and its adjacent frames $I_s \in \{t+1, t-1\}$ in a continuous sequence, the depth map is estimated by a depth net ϕ_D . Meanwhile, the camera pose transformation $T_{t \rightarrow s}$ between I_t and I_s is also predicted by a pose net ϕ_P . Subsequently, the pixel-wise correspondence regarding the target frame I_t and the source frame I_s can be constructed via the following equation:

$$p_s \sim K T_{t \rightarrow s} D_t(p_t) K^{-1} p_t, \quad (1)$$

where K is the camera intrinsic, p_t/p_s are corresponding pixels in I_t/I_s respectively. Based on Eq. 1, a reconstructed target frame $\hat{I}_{s \rightarrow t}$ is synthesized by bilinear sampling from adjacent source frames I_s .

As inaccurate depth will result in dissimilarity between I_t and $\hat{I}_{s \rightarrow t}$, a photometric reconstruction loss can be calculated to optimize the depth net ϕ_D and the pose net ϕ_P :

$$\mathcal{L}_{pe}^I = \gamma(1 - SSIM(\hat{I}_{s \rightarrow t}, I_t)) + (1 - \gamma) \|\hat{I}_{s \rightarrow t} - I_t\|_1, \quad (2)$$

where $SSIM$ measures the structure similarity between $\hat{I}_{s \rightarrow t}$ and I_t and γ sets to 0.85 following existing methods (Godard et al. 2019; Shao et al. 2022).

Moreover, an edge-aware smoothness loss is also adopted to encourage the sharpness of depth on the boundary of the object:

$$\mathcal{L}_{edge} = |\partial_x D_t| e^{-|\partial_x I_t|} + |\partial_y D_t| e^{-|\partial_y I_t|}. \quad (3)$$

3.2 Illumination-Reflectance-Depth Modeling for Endoscopic Images

Although this standard paradigm yields promising results in the natural image domain, its performance degrades in endoscopic scenes due to the uneven illumination.

While low-light enhancement can improve visibility (Reza 2004; Ma et al. 2022), they do not provide explicit guidance for learning illumination distributions. Moreover, approaches from other domains that handle lighting variations (Wang et al. 2021; Liu et al. 2021) rely on image collections captured under diverse lighting conditions, which are costly to acquire for endoscopic scenes. These limitations motivate our design of an illumination model tailored to endoscopy, aiming to improve depth estimation under uneven illumination without requiring specialized data.

Illumination-Reflectance-Depth (IRD) Model: According to the Retinex theory (Land and McCann 1971; Land 1977), an image can be represented by the element-wise Hadamard product of a reflectance map and an illumination map:

$$\hat{I}_t^r = R_t \cdot L_t, \quad (4)$$

where R is the reflectance map and L is the illumination map. Furthermore, endoscopic images are a type of near-field imaging and are often collected in a confined space. Consequently, anatomical structures farther from the camera tend to receive fewer light rays. Based on this observation, we consider that depth information also plays a key role in

the appearance of the images. Then, we can extend the Retienx theory (Land and McCann 1971; Land 1977) to the following formulation for the endoscopic images:

$$\hat{I}_t^r = R_t \cdot L_t \cdot e^{-\beta D_t}, \quad (5)$$

where β is the depth scaling factor that prevents D_t from over-dominant.

Another prominent characteristic of endoscopic images is their exclusive reliance on artificial illumination, with no contribution from ambient light sources. In the Lambertian model (Oren and Nayar 1995), the single source light can be formulated by a cosine function with an angular decay factor:

$$L_t = \cos^\theta(I_t, L_t^c) = \left(\frac{1}{1 + \|I_t - L_t^c\|_2} \right)^\theta, \quad (6)$$

where L_t^c is the center of the artificial light source, θ is the angular decay factor that controls the degradation speed of illumination.

IRD Decoupling: As can be seen in the green block of Fig. 2, we set two auxiliary networks ϕ_L and ϕ_R to predict the illumination and reflectance map.

$$\begin{aligned} R_t &= \phi_R(I_t) \\ (\theta, x_0, y_0) &= \phi_L(I_t). \end{aligned} \quad (7)$$

Notably, the illumination net estimates the parameters of the illumination map, and (x_0, y_0) are the pixel coordinates of the center L_t^c . Combined with the depth map, we can decouple the endoscopic images from the IRD modeling and recover the texture details of low-light regions in the endoscopic images.

3.3 Joint Optimization of IRD Decoupling and Depth Estimation

Based on the IRD model, we propose a self-supervised joint optimization framework for monocular depth estimation and IRD decoupling.

$$\begin{cases} \hat{I}_t^r &= R_t \cdot L_t \cdot e^{-\beta D_t} \\ R_t &\stackrel{\text{sampling}}{\leftarrow} (R_s, p_s), p_s \sim K T_{t \rightarrow s} D_t(p_t) K^{-1} p_t \end{cases} \quad (8)$$

As shown in formula Eq. 8, the benefits are twofold. On one hand, as the depth map is part of the IRD model, the more precise the depth prediction is, the better the reflectance and illumination map can be learned. On the other hand, since the decoupled reflectance recovers the intrinsic appearance in low-light regions, it can be used for view synthesis between adjacent frames and target frames. By calculating photometric loss on the synthesized reflectance and the true reflectance, the depth net and the pose net receive the correct gradient backpropagation from the low-light regions, thus having better performance. Additionally, the illumination map can be applied to the original RGB images, serving as a simple yet powerful augmentation to strengthen the generalized ability of the depth net.

IRD decoupling branch: We propose two losses for training the IRD decoupling branch. With each component of the IRD model estimated by the depth net and two auxiliary networks, the reconstructed target frame \hat{I}_t^r is acquired by element-wise multiplication. Subsequently, a reconstruction loss is calculated between the \hat{I}_t^r and the real target frame I_t :

$$\begin{aligned} \hat{I}_t^r &= \phi_R(I_t) \cdot L_t^c(\phi_L(I_t)) \cdot e^{\beta D_t} \\ \mathcal{L}_{rec} &= \gamma(1 - SSIM(\hat{I}_t^r, I_t)) + (1 - \gamma)\|\hat{I}_t^r - I_t\|_1 \end{aligned} \quad (9)$$

where $L_t^c(\cdot)$ builds the illumination map from the estimated (x_0, y_0) , and θ parameters.

However, simply relying on reconstruction loss may easily fall into the trivial solution, where all predicted illumination maps remain consistent and the reflectances are the same as the target frame I_t . As a solution, the illumination net is regularized by comparing the intensity ratio between the target frame I_t and the illumination map L_t . To be more concrete, with p_{95} and p_{20} being the percentile intensity value, we hypothesize that a correct illumination map should have a ratio between high and low intensity similar to that of the target frame I_t .

$$\mathcal{L}_{ratio} = \left\| \frac{p_{95}(I_t)}{p_{20}(I_t)} - \frac{p_{95}(L_t)}{p_{20}(L_t)} \right\|_1 \quad (10)$$

By these two loss functions, we can prevent the trivial solution and decouple well-illuminated reflectance in a self-supervised manner.

Depth estimation branch: Illustrated in the gray block of Fig. 2, we consider how the decoupled IRD component can be used to enhance the performance of the depth estimation model. As mentioned previously, we notice that the reflectance recovers the texture details in the dark areas. Thereafter, we can synthesize target reflectance $\hat{R}_{s \rightarrow t}$ from the source reflectance R_s , and calculate the photometric loss similar to Eq. 2.

$$\mathcal{L}_{pe}^R = \gamma(1 - SSIM(\hat{R}_{s \rightarrow t}, R_t)) + (1 - \gamma)\|\hat{R}_{s \rightarrow t} - R_t\|_1, \quad (11)$$

By adding this auxiliary loss, the depth and pose net will receive correct backpropagation gradients in the low-light regions. In addition, to further enhance the robustness of the depth estimation model against uneven illumination, we directly combine the predicted depth D_t and illumination maps L_t with the original target frame to generate a degraded image \hat{I}_t^{dg} .

$$\hat{I}_t^{dg} = I_t \cdot L_t \cdot e^{-\beta D_t} \quad (12)$$

We hypothesize that if the depth net is robust enough to the uneven illumination, the depth output from I_t and \hat{I}_t^{dg} should be the same. Therefore, a degradation consistency loss is introduced:

$$\mathcal{L}_{dg} = \|\phi_D(I_t) - \phi_D(\hat{I}_t^{dg})\|_1 \quad (13)$$

Overall Loss: Finally, we obtain the overall losses for networks of both the depth estimation branch and IRD decoupling branch.

$$\begin{cases} \mathcal{L}_{\phi_D} = \mathcal{L}_{pe}^I + \mathcal{L}_{pe}^R + \mathcal{L}_{dg} \\ \mathcal{L}_{\phi_P} = \mathcal{L}_{pe}^I + \mathcal{L}_{pe}^R \\ \mathcal{L}_{\phi_R} = \mathcal{L}_{rec} \\ \mathcal{L}_{\phi_L} = \mathcal{L}_{rec} + \mathcal{L}_{ratio} \end{cases} \quad (14)$$

4 Experiments

4.1 Datasets

We evaluate our method using two public endoscopic datasets for both comparative and ablative studies.

SCARED: The SCARED (Allan et al. 2021) dataset consists of stereo laparoscopic videos acquired from fresh porcine cadaver abdominal anatomy, with 35 stereo videos recorded across 9 distinct scenes. Adhering to previous work (Huang et al. 2022), we select 15351 frames for training, 1705 for validation, and 90 keyframes with accurate depth labels acquired using structured light for testing.

Hamlyn: The Hamlyn (Recasens et al. 2021) dataset contains in vivo stereo endoscopic videos captured during various surgical procedures. It poses greater challenges due to low-light conditions and significant inter-frame motion. In line with prior protocols (Cui et al. 2024; Recasens et al. 2021), we use all 21 videos for generalization evaluation.

4.2 Implementations

We implement our method with Pytorch on a single NVIDIA RTX A6000. Adam (Kingma and Ba 2015) optimizer is used with $\beta_1=0.9$, $\beta_2=0.99$. The initial learning rate is set to $1e-4$ and scales down by 0.3 every 10 epochs. The total training epoch is 30 with a batch size of 12. Cohere to previous methods (Godard et al. 2019; Cui et al. 2024), preprocessing contains resizing the image to 320×256 , randomly flipping, and color jittering. We select two different base models for a comprehensive evaluation: Monodepth2 (Godard et al. 2019) with ResNet-18, and DepthAnything (Yang et al. 2024) with Vit-B. In addition, the implementation of illumination net ϕ_L and the pose net ϕ_P is a simple ResNet-18 regression network, while the reflectance network ϕ_R is adopted from the RetinexNet (Chen et al. 2018).

4.3 Metrics

We adopt commonly-used depth estimation indicators, including Abs Rel, Sq Rel, RMSE, RMSE log errors, and the $\delta < 1.25^1$, $\delta < 1.25^2$, $\delta < 1.25^3$ accuracy metrics, to comprehensively assess model performance. See the supplement for more details.

4.4 Effectiveness of Illumination Decoupling

While low-light enhancement methods from natural image processing offer potential for endoscopic illumination, our analysis shows their limitations and highlights the need for accurate illumination modeling in endoscopic scenes.

Two traditional statistics-based methods(CLAHE (Reza 2004) and LIME (Guo, Li, and Ling 2016)) and one unsupervised learning-based method(SCI (Ma et al. 2022)) are

Methods	Abs Rel↓	Sq Rel↓	$\delta < 1.25^1 \uparrow$	$\delta < 1.25^2 \uparrow$
Monodepth2(-)	0.075	0.799	0.952	0.990
CLAHE(S)	0.072	0.787	0.958	0.991
LIME(S)	0.077	0.819	0.949	0.992
SCI(D)	0.073	0.790	0.954	0.992
DeLightMono(D)	0.065	0.655	0.961	0.993

S: statistics-based method, D: learning-based method. \uparrow : the higher the better, \downarrow : the lower the better. The best results are reported in **bold**.

Table 1: Performance comparison with image enhancement methods. All models are based on Monodepth2 and evaluated on the SCARED dataset.

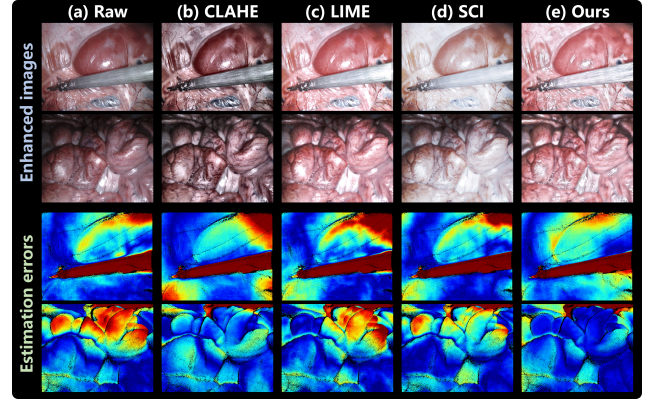


Figure 3: Enhancement results and error maps on the SCARED dataset.

selected for comparison. All these methods serve as a pre-processing strategy during training and inference. As Tab. 1 demonstrates, compared to the baseline Monodepth2 (Godard et al. 2019), the statistic-based method CLAHE (Reza 2004) and the learning-based method SCI (Ma et al. 2022) both slightly improve the performance. However, such amelioration is minor, and for LIME (Guo, Li, and Ling 2016), the situation is even deteriorating.

A more intuitive illustration is shown in Fig. 3 with enhancement results and error maps. From the top cases, CLAHE (Reza 2004) amplified the texture details in the image, while SCI (Ma et al. 2022) and LIME (Guo, Li, and Ling 2016) lit up the illumination level globally. However, row 1 shows that neither the light level nor the texture details are recovered in the dark areas of these methods' results, such as the top-right corner. Additionally, SCI (Ma et al. 2022), trained on the SCARED (Allan et al. 2021) before being applied, introduces color distortion. Consequently, estimation error does not decline and even increases (row 3, columns b and d). For another case, as in rows 2 and 4, LIME (Guo, Li, and Ling 2016) successfully recovers the light level and details at the top of the image. Nevertheless, the performance improvement is also limited.

These findings imply the necessity of modeling the illumination and utilizing the illumination as a guide during the training of the depth estimation network. In contrast, our proposed framework jointly optimizes the IRD decou-

Method	Year	Backbone	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25^1\uparrow$	$\delta < 1.25^2\uparrow$	$\delta < 1.25^3\uparrow$
Monodepth2	2019	ResNet-18	0.075	0.794	6.213	0.099	0.952	0.991	0.998
HR-Depth	2021	ResNet-18	0.069	0.731	5.777	0.093	0.959	0.993	0.997
M3Depth	2022	ResNet-18	0.072	0.764	6.178	0.098	0.961	0.992	0.996
AF-SfMLearner	2022	ResNet-18	0.075	0.798	6.212	0.100	0.953	0.991	0.997
IID-SfMLearner	2024	ResNet-18	0.072	0.789	5.938	0.096	0.961	0.991	0.996
DeLightMono(Ours)	-	ResNet-18	0.065	0.655	5.589	0.089	0.961	0.993	0.998
DA(zero-shot)	2024	ViT-B	0.197	4.685	14.60	0.245	0.729	0.903	0.965
DA(finetune)	2024	ViT-B	0.065	0.563	5.640	0.088	0.962	0.997	0.999
MonoViT	2022	ViT-B	0.065	0.642	5.492	0.088	0.961	0.994	0.998
EndoDAC	2024	ViT-B	0.064	0.607	5.715	0.088	0.963	0.995	0.998
DeLightMono(Ours)	-	ViT-B	0.059	0.496	5.214	0.081	0.969	0.997	1.000

↑: the higher the better, ↓: the lower the better. The best results are reported in **bold**.

Table 2: Quantitative results on the SCARED dataset.

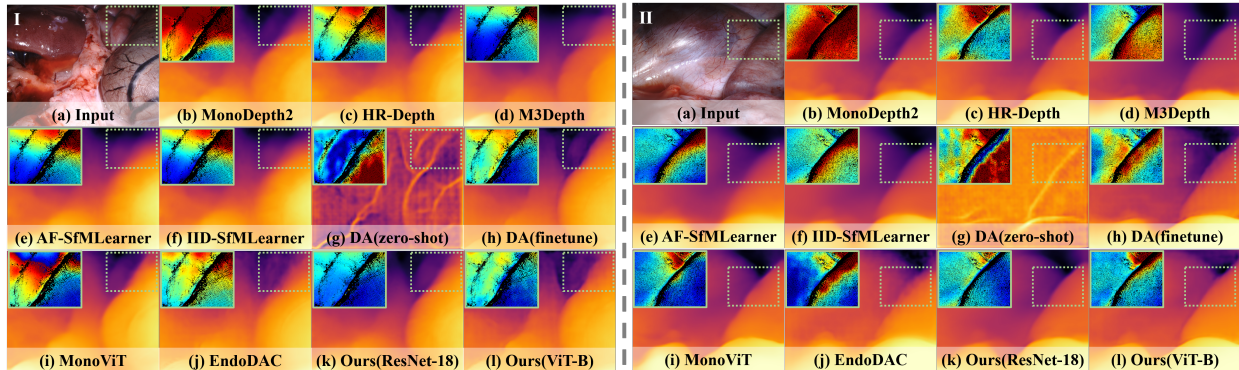


Figure 4: Qualitative comparisons on the SCARED dataset with error maps of the highlighted regions.

pling and depth estimation task, resulting in restored details and performance enhancement in dark areas, as evidenced in rows 3 and 4 of Fig. 3.

4.5 Compare with SOTA methods

Comparison on SCARED dataset: Tab. 2 shows a comparison with the state-of-the-art methods both from the natural-image and endoscopic domains on the SCARED (Allan et al. 2021) dataset. Monodepth2 (Godard et al. 2019), as the baseline method, suffers from performance degradation due to the uneven illumination, while HR-Depth (Lyu et al. 2021) redesigns the skip-connection to fuse high-resolution features, thereby achieving better results. M3depth (Huang et al. 2022) focuses on 3D geometric consistency between frames, while Af-SfMLearner (Shao et al. 2022) and IID-SfMLearner (Li et al. 2024) are solutions for varying specular reflection between adjacent frames. As they do not account for the spatially uneven illumination, their improvements are also marginal. The depth foundation model Depth anything (Yang et al. 2024) exhibits robust zero-shot ability across various scenes. Nevertheless, its performance declined significantly, which shows the gaps between natural images and endoscopic images. After finetuning, it shows similar performance compared to MonoViT (Zhao et al. 2022) and EndoDAC (Cui et al. 2024).

Our DeLightMono(ViT-B) achieves the best performance across all metrics. What is more noteworthy is that our DeLightMono(ResNet-18) achieves comparative results

with ViT-B-based methods, which validates the feasibility of our illumination model and the joint optimization framework.

Furthermore, we visualize the depth outputs and their error maps in low-light regions in Fig. 4. As the angle between the regions and the light source increases, their local illumination diminishes significantly, leading to a textureless appearance. Since other methods do not take the varying illumination of endoscopic images into account, their error rates grow correspondingly. On the contrary, benefitting from the reflectance and our joint optimization framework, our methods not only maintain superior results but also have clearer depth prediction in these regions.

Comparison on Hamlyn dataset: Fig. 5 and Tab. 3 shows the generalization comparisons on the Hamlyn (Recasens et al. 2021) dataset. As observed, Hamlyn (Recasens et al. 2021) presents a significant difference in appearance compared to SCARED (Allan et al. 2021), especially the illumination levels. Comparison models with the ResNet-18 backbone struggle to recognize the shapes of tissues and surgical instruments in the depth maps. Conversely, our model still roughly captures edges and the relative depth. Benefiting from contextual understanding of the vision transformer, ViT-B-based models offer substantial alleviation of the situation. Combined with our proposed method, the depth net accurately delineates instrument shapes and makes precise depth predictions. Furthermore, the quantitative results presented in the data table align well with the visualized out-

Method	Year	Backbone	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25^1\uparrow$	$\delta < 1.25^2\uparrow$	$\delta < 1.25^3\uparrow$
Monodepth2	2019	ResNet-18	0.165	4.586	16.973	0.218	0.763	0.919	0.974
HR-Depth	2021	ResNet-18	0.162	4.416	16.713	0.214	0.767	0.922	0.976
M3Depth	2022	ResNet-18	0.161	4.377	16.497	0.213	0.768	0.921	0.976
AF-SfMLearner	2022	ResNet-18	0.179	5.665	17.693	0.226	0.759	0.913	0.968
IID-SfMLearner	2024	ResNet-18	0.168	4.787	16.944	0.217	0.767	0.921	0.974
DeLightMono(Ours)	-	ResNet-18	0.158	4.095	16.127	0.208	0.773	0.929	0.980
DA(zero-shot)	2024	ViT-B	0.174	4.473	18.307	0.238	0.718	0.912	0.975
DA(finetune)	2024	ViT-B	0.146	3.546	15.217	0.198	0.789	0.936	0.984
MonoViT	2022	ViT-B	0.158	4.215	16.403	0.211	0.776	0.927	0.978
EndoDAC	2024	ViT-B	0.146	3.861	15.058	0.194	0.802	0.937	0.980
DeLightMono(Ours)	-	ViT-B	0.144	3.648	14.989	0.196	0.799	0.937	0.981

↑: the higher the better, ↓: the lower the better. The best results are reported in **bold**.

Table 3: Generalization results on the Hamlyn dataset. All methods are trained on the SCARED dataset.

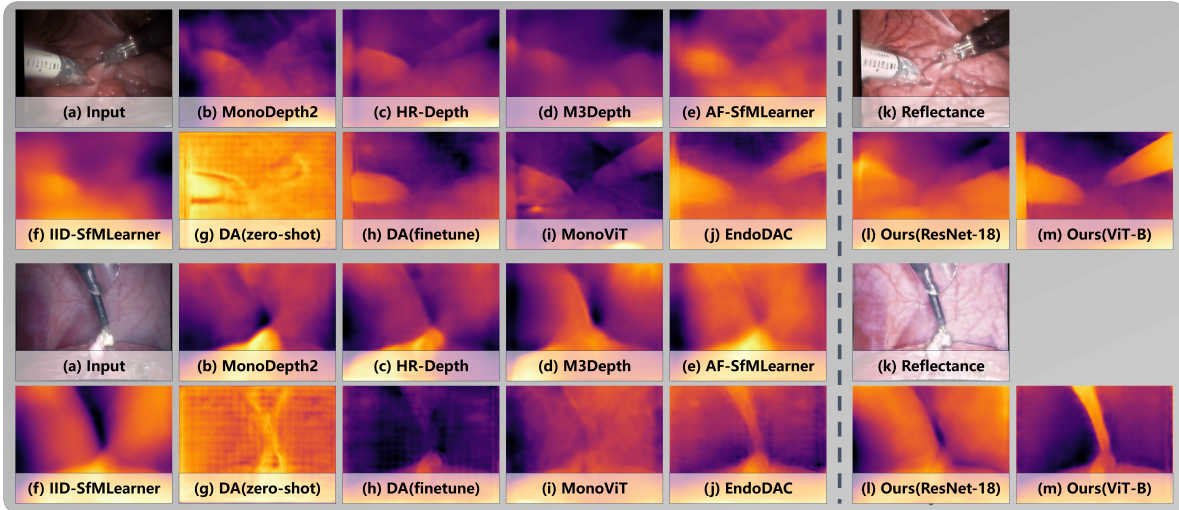


Figure 5: Qualitative comparisons on the Hamlyn dataset. Additionally, reflectances from the proposed method are provided.

puts, and our models achieve leading performance across most of the evaluation metrics.

4.6 Ablation studies

To validate the contribution of each component to our proposed methods, ablation studies are conducted on SCARED (Allan et al. 2021) by removing each of them in turn. The results are reported in Tab. 4.

Comparing rows 2 and 6, the performance drops if we remove the reflectance-based photometric loss term. The result shows that the recovered illumination and texture in low-light regions act as key factors in enhancing the performance of the depth estimation model. Meanwhile, a similar performance deterioration takes place when comparing rows 3 and 6, further demonstrating that our degradation consistency loss contributes to enhancing the model’s robustness against illumination variations in endoscopic images.

We observe a more drastic decline in estimation accuracy once the \mathcal{L}_{ratio} or \mathcal{L}_{rec} is removed. More specifically, the illumination map becomes a uniform distribution on the model w/o \mathcal{L}_{ratio} , while the reflectance loses most of the semantic information if the \mathcal{L}_{rec} is left. This indicates that both \mathcal{L}_{ratio} and \mathcal{L}_{rec} serve as the key regularization terms

for proper IRD decoupling.

Depth			Decoupling		Abs Rel↓	Sq Rel↓	RMSE↓
L_{pe}^I	L_{pe}^R	L_{dg}	L_{rec}	L_{ratio}			
✓					0.075	0.794	6.213
✓		✓	✓	✓	0.067	0.682	5.631
✓	✓		✓	✓	0.067	0.692	5.629
✓	✓	✓		✓	0.069	0.712	5.812
✓	✓	✓	✓		0.069	0.744	5.860
✓	✓	✓	✓	✓	0.065	0.655	5.589

Table 4: Ablations on the SCARED dataset.

5 Conclusion

In this work, we propose DeLightMono, a self-supervised depth estimation algorithm that addresses uneven illumination in endoscopic scenarios through image decoupling. The method decomposes an endoscopic image with the illumination-reflectance-depth modeling, and harnesses these components to strengthen the depth net against uneven illumination via a joint optimization framework. The results of comprehensive comparison experiments and ablation studies demonstrate the advancing performance of the proposed method.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62401246), and Shenzhen Science and Technology Program (JCYJ20250604185805008, JCYJ20240813095112017)

References

- Allan, M.; McLeod, A. J.; Wang, C.; Rosenthal, J.; Hu, Z.; Gard, N.; Eisert, P.; Fu, K. X.; Zeffiro, T.; Xia, W.; Zhu, Z.; Luo, H.; Jia, F.; Zhang, X.; Li, X.; Sharan, L.; Kurmann, T.; Schmid, S.; Sznitman, R.; Psychogyios, D.; Azizian, M.; Stoyanov, D.; Maier-Hein, L.; and Speidel, S. 2021. Stereo Correspondence and Reconstruction of Endoscopic Data Challenge. *CoRR*, abs/2101.01133.
- Amitai, S.; Klein, I.; and Treibitz, T. 2023. Self-supervised monocular depth underwater. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 1098–1104. IEEE.
- Budd, C.; and Vercauteren, T. 2024. Transferring Relative Monocular Depth to Surgical Vision with Temporal Consistency. In *27th MICCAI*, 692–702. Springer.
- Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; and Zhang, Y. 2023. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12504–12513.
- Chen, W.; Wenjing, W.; Wenhao, Y.; and Jiaying, L. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *British Machine Vision Conference*. British Machine Vision Association.
- Cui, B.; Islam, M.; Bai, L.; Wang, A.; and Ren, H. 2024. Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In *27th MICCAI*, 208–218. Springer.
- Eigen, D.; Puhersch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.
- Engel, J.; Schöps, T.; and Cremers, D. 2014. LSD-SLAM: Large-scale direct monocular SLAM. In *European conference on computer vision*, 834–849. Springer.
- Enqvist, O.; Kahl, F.; and Olsson, C. 2011. Non-sequential structure from motion. In *2011 ICCV Workshops*, 264–271. IEEE.
- Fan, W.; Jiang, W.; Shi, H.; Zeng, H.-Q.; Chen, Y.; and Luo, X. 2024. Triple-Supervised Convolutional Transformer Aggregation for Robust Monocular Endoscopic Dense Depth Estimation. *IEEE Transactions on Medical Robotics and Bionics*, 6(3): 1017–1029.
- Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep ordinal regression network for monocular depth estimation. In *CVPR 2018*, 2002–2011.
- Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Un-supervised monocular depth estimation with left-right consistency. In *CVPR 2017*, 270–279.
- Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3828–3838.
- Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1780–1789.
- Guo, X.; Li, Y.; and Ling, H. 2016. LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2): 982–993.
- 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*.
- Huang, B.; Zheng, J.-Q.; Nguyen, A.; Xu, C.; Gkouzionis, I.; Vyas, K.; Tuch, D.; Giannarou, S.; and Elson, D. S. 2022. Self-supervised Depth Estimation in Laparoscopic Image Using 3D Geometric Consistency. In *25th MICCAI*, 13–22. Springer.
- Karsch, K.; Liu, C.; and Kang, S. B. 2014. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence*, 36(11): 2144–2158.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Klingner, M.; Termöhlen, J.-A.; Mikolajczyk, J.; and Fingscheidt, T. 2020. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European conference on computer vision*, 582–600. Springer.
- Land, E. H. 1977. The retinex theory of color vision. *Scientific american*, 237(6): 108–129.
- Land, E. H.; and McCann, J. J. 1971. Lightness and retinex theory. *Journal of the Optical society of America*, 61(1): 1–11.
- Li, B.; Liu, B.; Zhu, M.; Luo, X.; and Zhou, F. 2024. Image Intrinsic-Based Unsupervised Monocular Depth Estimation in Endoscopy. *IEEE Journal of Biomedical and Health Informatics*.
- Li, B.; Shen, C.; Dai, Y.; Van Den Hengel, A.; and He, M. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR 2015*, 1119–1127.
- Li, W.; Hayashi, Y.; Oda, M.; Kitasaka, T.; Misawa, K.; and Mori, K. 2022. Geometric constraints for self-supervised monocular depth estimation on laparoscopic images with dual-task consistency. In *25th MICCAI*, 467–477. Springer.
- Liu, L.; Song, X.; Wang, M.; Liu, Y.; and Zhang, L. 2021. Self-supervised Monocular Depth Estimation for All Day Images using Domain Separation. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, 12737–12746.
- Lu, Y.; Wei, R.; Li, B.; Chen, W.; Zhou, J.; Dou, Q.; Sun, D.; and Liu, Y.-h. 2023. Autonomous intelligent navigation for flexible endoscopy using monocular depth guidance and 3-D shape planning. In *2023 IEEE international conference on robotics and automation (ICRA)*, 1–7. IEEE.
- Lyu, X.; Liu, L.; Wang, M.; Kong, X.; Liu, L.; Liu, Y.; Chen, X.; and Yuan, Y. 2021. Hr-depth: High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2294–2301.
- Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5637–5646.
- Metzger, R.; Suppa, P.; Li, Z.; and Vemuri, A. 2024. Augmented reality navigation systems in endoscopy. *Frontiers in Gastroenterology*, 3: 1345466.
- Oren, M.; and Nayar, S. K. 1995. Generalization of the Lambertian model and implications for machine vision. *International Journal of Computer Vision*, 14(3): 227–251.
- Ozyoruk, K. B.; Gokceler, G. I.; Bobrow, T. L.; Coskun, G.; Incetan, K.; Almalioglu, Y.; Mahmood, F.; Curto, E.; Perdigo, 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.
- Pizer, S. M.; Amburn, E. P.; Austin, J. D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J. B.; and Zuiderveld, K. 1987. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3): 355–368.
- Poggi, M.; Aleotti, F.; Tosi, F.; and Mattoccia, S. 2020. On the uncertainty of self-supervised monocular depth estimation. In *CVPR 2020*, 3227–3237.
- Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637.
- Recasens, D.; Lamarca, J.; Fàcil, 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.
- Reza, A. M. 2004. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38: 35–44.
- Shao, S.; Pei, Z.; Chen, W.; Zhu, W.; Wu, X.; Sun, D.; and Zhang, B. 2022. Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue. *Medical image analysis*, 77: 102338.
- Vankadari, M.; Golodetz, S.; Garg, S.; Shin, S.; Markham, A.; and Trigoni, N. 2023. When the Sun Goes Down: Repairing Photometric Losses for All-Day Depth Estimation. In Liu, K.; Kulic, D.; and Ichnowski, J., eds., *Proceedings of The 6th CoRL*, volume 205 of *Proceedings of Machine Learning Research*, 1992–2003. PMLR.
- Wang, K.; Zhang, Z.; Yan, Z.; Li, X.; Xu, B.; Li, J.; and Yang, J. 2021. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16055–16064.
- Wu, Y.; Pan, C.; Wang, G.; Yang, Y.; Wei, J.; Li, C.; and Shen, H. T. 2023. Learning semantic-aware knowledge guidance for low-light image enhancement. In *CVPR 2023*, 1662–1671.
- Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.
- Yue, G.; Gao, J.; Cong, R.; Zhou, T.; Li, L.; and Wang, T. 2023. Deep pyramid network for low-light endoscopic image enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5): 3834–3845.
- Yue, H.; and Gu, Y. 2023. Tcl: Triplet consistent learning for odometry estimation of monocular endoscope. In *26th MICCAI*, 144–153. Springer.
- Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond brightening low-light images. *International Journal of Computer Vision*, 129: 1013–1037.
- 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 *2022 3DV*, 668–678. IEEE.
- Zheng, Y.; Zhong, C.; Li, P.; Gao, H.-a.; Zheng, Y.; Jin, B.; Wang, L.; Zhao, H.; Zhou, G.; Zhang, Q.; et al. 2023. Steps: Joint self-supervised nighttime image enhancement and depth estimation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 4916–4923. IEEE.
- Zhou, K.; Bian, J.-W.; Zheng, J.-Q.; Zhong, J.; Xie, Q.; Trigoni, N.; and Markham, A. 2025. Manydepth2: Motion-aware self-supervised monocular depth estimation in dynamic scenes. *IEEE Robotics and Automation Letters*.
- Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *CVPR 2017*, 1851–1858.
- Zhu, J.; Lyu, L.; Xu, Y.; Liang, H.; Zhang, X.; Ding, H.; and Wu, Z. 2021. Intelligent soft surgical robots for next-generation minimally invasive surgery. *Advanced Intelligent Systems*, 3(5): 2100011.