Medical Image Analysis (2024)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

MonoPCC: Photometric-invariant Cycle Constraint for Monocular Depth Estimation of Endoscopic Images

Zhiwei Wang[a,1], Ying Zhou[a,1], Shiquan He[a], Ting Li[b], Fan Huang[b], Qiang Ding[c], Xinxia Feng[c], Mei Liu[c], Qiang Li[a,*]

[a] *Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, 430074, China.*
[b] *Wuhan United Imaging Healthcare Surgical Technology Co., Ltd., Wuhan, 430074, China.*
[c] *Department of Gastroenterology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430074, China*

ARTICLE INFO

Article history:
Received None
Received in final form None
Accepted None
Available online None

Keywords: Self-supervised learning, Monocular depth estimation, Photometric constraint, Brightness robustness

ABSTRACT

Photometric constraint is indispensable for self-supervised monocular depth estimation. It involves warping a source image onto a target view using estimated depth&pose, and then minimizing the difference between the warped and target images. However, the endoscopic built-in light causes significant brightness fluctuations, and thus makes the photometric constraint unreliable. Previous efforts only mitigate this relying on extra models to calibrate image brightness. In this paper, we propose MonoPCC to address the brightness inconsistency radically by reshaping the photometric constraint into a cycle form. Instead of only warping the source image, MonoPCC constructs a closed loop consisting of two opposite forward-backward warping paths: from target to source and then back to target. Thus, the target image finally receives an image cycle-warped from itself, which naturally makes the constraint invariant to brightness changes. Moreover, MonoPCC transplants the source image's phase-frequency into the intermediate warped image to avoid structure lost, and also stabilizes the training via an exponential moving average (EMA) strategy to avoid frequent changes in the forward warping. The comprehensive and extensive experimental results on four endoscopic datasets demonstrate that our proposed MonoPCC shows a great robustness to the brightness inconsistency, and exceeds other state-of-the-arts by reducing the absolute relative error by at least 7.27%, 9.38%, 9.90% and 3.17%, respectively.

1. Introduction

Monocular endoscope is the key medical imaging tool for gastrointestinal diagnosis and surgery, but often provides a narrow field of view (FOV). 3D scene reconstruction helps enlarge the FOV, and also enables more advanced applications like surgical navigation by registration with preoperative computed tomography (CT). Depth estimation of monocular endoscopic images is prerequisite for reconstructing 3D structures, but extremely challenging due to absence of ground-truth (GT) depth labels.

The typical solution of monocular depth estimation relies on self-supervised learning, where the core idea is photometric constraint between real and warped images. Specifically, two convolutional neural networks (CNNs) have to be built; one is called DepthNet and the other is PoseNet. The two CNNs estimate a depth map of each image and camera pose changes of every two adjacent images, based on which a source frame in

*Corresponding author.
e-mail: liqiang8@hust.edu.cn (Qiang Li)
[1]Equal contribution.

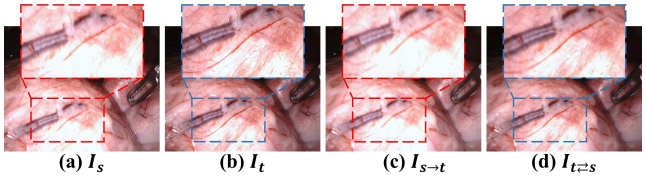

Fig. 1. (a)-(b) are the source I_s and target I_t frames. (c) is the warped image from the source to target. (d) is the cycle-warped image along the target-source-target path for reliable photometric constraint. Box contour colors distinguish different brightness patterns.

an endoscopic video can be projected into 3D space and warped onto a target view of the other frame. DepthNet and PoseNet are jointly optimized to minimize a photometric loss, which is essentially the pixel-to-pixel difference between the warped and target images.

However, the light source is fixed to the endoscope and moves along with the camera, resulting in significant brightness fluctuations between the source and target frames. Such problem can also be worsened by non-Lambertian reflection due to the close-up observation, as evidenced in Fig. 1(a)-(b). Consequently, between the target and warped source images, the brightness difference dominates as shown in Fig. 1(b)-(c), and thus misguides the photometric constraint in the self-supervised learning.

Many efforts have been paid to enhance the reliability of photometric constraint under the fluctuated brightness. An intuitive solution is to calibrate the brightness in endoscopic video frames beforehand, using either a linear intensity transformation (Ozyoruk et al., 2021) or a trained appearance flow model (Shao et al., 2022). However, the former only addresses the global brightness inconsistency, and the latter increases the training difficulty due to introduced burdensome computation. Moreover, the reliability of appearance flow model is also not always guaranteed due to the weak self-supervision, which leads to a risk of wrongly modifying areas unrelated to brightness changes.

In this paper, we aim to address the bottleneck of brightness inconsistency without relying on any auxiliary model. Our motivation stems from a recent method named TC-Depth (Ruhkamp et al., 2021), which introduced a cycle warping originally for solving the occlusion issue. TC-Depth warps a target image to source and then warps back to itself to identify every occluded pixel, since they assume the occluded pixel can not come back to the original position precisely. We find that such cycle warping can naturally overcome the brightness inconsistency, and yield a more reliable warped image compared to only warping from source to target, as shown in Fig. 1(b)-(d). However, directly applying the cycle warping often fails in photometric constraint because (1) the twice bilinear interpolation of cycle warping blurs the image too much, and (2) the networks of depth&pose estimation are learned actively, which makes the intermediate warping unstable and the convergence difficult.

In view of the above analysis, we propose **Mono**cular depth estimation based on **P**hotometric-invariant **C**ycle **C**onstraint (MonoPCC), which adopts the idea of cycle warping but significantly reshapes it to enable the photometric constraint invariant to inconsistent brightness. Specifically, MonoPCC starts from the target image, and follows a closed loop path (target-source-target) to obtain a cycle-warped target image, which inherits consistent brightness from the original target image. To make such cycle warping effective in the photometric constraint, MonoPCC employs a learning-free structure transplant module (STM) based on Fast Fourier Transform (FFT) to minimize the negative impact of the blurring effect. STM restores the lost structure details in the intermediate warped image by 'borrowing' the phase-frequency part from the source image. Moreover, instead of sharing the network weights in both target-source and source-target warping paths, MonoPCC bridges the two paths using an exponential moving average (EMA) strategy to stabilize the intermediate results in the first path.

In summary, our main contributions are as follows:

1. We propose MonoPCC, which eliminates the inconsistent brightness inducing misguidance in the self-supervised learning by simply adopting a cycle-form warping to render the photometric constraint invariant to the brightness changes.
2. We introduce two enabling techniques, i.e., structure transplant module (STM) and an EMA-based stable training. STM restores the lost image details caused by interpolation, and EMA stabilizes the forward warping. These together guarantee an effective training of MonoPCC under the cycle-form warping.
3. We conduct comprehensive and extensive experiments on four public endoscopic datasets, i.e., SCARED (Allan et al., 2021), SimCol3D (Rau et al., 2023), SERV-CT (Edwards et al., 2022), and Hamlyn (Mountney et al., 2010; Stoyanov et al., 2010; Pratt et al., 2010) and a public natural dataset, i.e., KITTI (Geiger et al., 2012). The comparison results with eight state-of-the-art methods demonstrate the superiority of MonoPCC by decreasing the absolute relative error by at least 7.27%, 9.38%, 9.90% and 3.17% on the four endoscopic datasets, respectively, and its strong ability to resist inconsistent brightness in the training. Moreover, the comparison results on KITTI further verify the competitiveness of MonoPCC even for the natural scenario, where the brightness changes are often not that significant.

2. Related Work

2.1. Self-Supervised Depth Estimation

Self-supervised methods (Zhou et al., 2017; Yin and Shi, 2018; Bian et al., 2019; Godard et al., 2019; Watson et al., 2021) of monocular depth estimation mostly, if not all, rely on the photometric constraint to train two networks, named DepthNet and PoseNet. For example, as a pioneering work, SfMLearner (Zhou et al., 2017) estimated depth maps and camera poses to synthesize a fake target image by warping another image from the source view, and calculated a reconstruction loss between the warped and target images as the photometric constraint.

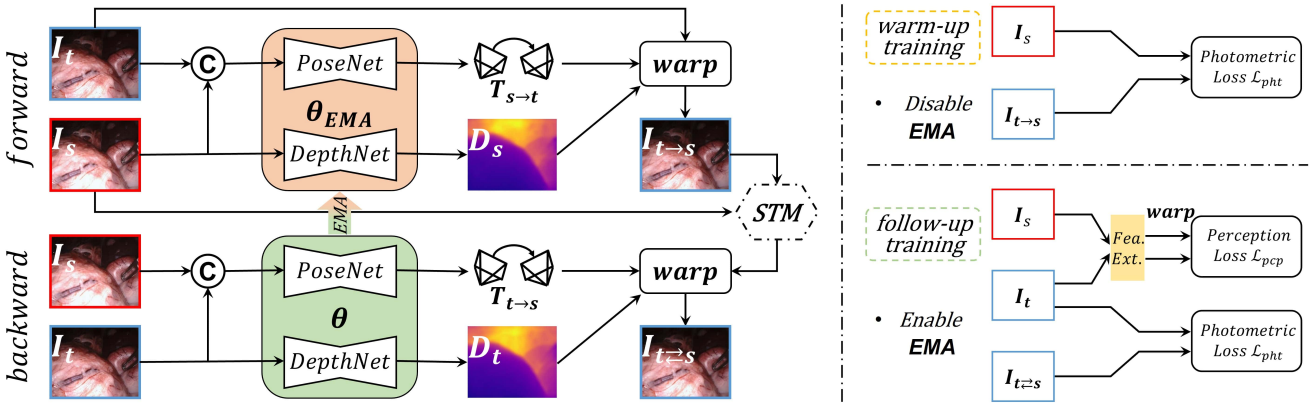

Fig. 2. The training pipeline of MonoPCC, which consists of forward and backward cascaded warping paths bridged by two enabling techniques, i.e., structure transplant module (STM) and exponential moving average (EMA). The training has two phases, i.e., warm-up to initialize the network weights for reasonable forward warping, and follow-up to resist the brightness changes. Different box contour colors code different brightness patterns. © means concatenation.

For endoscopic cases, M3Depth (Huang et al., 2022) enforced not only the photometric constraint but also a 3D geometric structural consistency by Mask-ICP. For improving pose estimation in laparoscopic procedures, Li et al. (2022) designed a Siamese pose estimation scheme and also constructed dual-task consistency by additionally predicting scene coordinates. These methods share a common flaw: brightness fluctuations caused by the moving light source in the endoscope, significantly reduce the effectiveness of photometric constraint in self-supervised learning, degrading the performance consequently.

2.2. Brightness Inconsistency

Recently, several efforts have been made to mitigate brightness inconsistency in endoscopic cases, the common solution is to calibrate the brightness inconsistency using either preprocessing or extra learned models. For example, Endo-SfMLearner (Ozyoruk et al., 2021) calculated the mean values and standard deviation of images, and then linearly aligned the brightness of the warped image with that of the target one. However, the real brightness fluctuations are actually local and non-linear, and are hardly calibrated by a simple linear transformation. Recently, AF-SfMLearner (Shao et al., 2022) tried to relax the requirement of brightness consistency by learning an appearance flow model, which estimates pixel-wise brightness offset between adjacent frames. Nevertheless, effectiveness relies on the quality of the learned appearance flow model, and the risk of wrongly modifying the image content occurs accompanying the model error. Also, introducing an extra model dramatically increases the training time and difficulty in self-supervised learning.

3. Method

Fig. 2 illustrates the pipeline of MonoPCC, consisting of both forward and backward warping paths in the training phase. We first explain how to warp images for self-supervised learning in Sec. 3.1, and then detail the photometric-invariant principle of MonoPCC in Sec. 3.2, as well as its two key enabling techniques in Sec. 3.3 and Sec. 3.4, i.e., structure transplant module (STM) for avoiding detail lost and EMA between two paths for stabilizing the training.

3.1. Warping across Views for Self-Supervision

To warp a source image I_s to a target view I_t, DepthNet and PoseNet first estimate the target depth map D_t, and the camera pose changing from target to source $T_{t\to s}$, respectively.

The DepthNet is typically an encoder-decoder, which inputs a single endoscopic image and outputs an aligned depth map. In this work, we use MonoViT (Zhao et al., 2022) as our backbone DepthNet. The PoseNet contains a lightweight ResNet18-based (He et al., 2016) encoder with four convolutional layers, which inputs a concatenated two images and predicts the pose change from the image in the tail channels to that in the front channels.

Using D_t and $T_{t\to s}$, the warping can be performed from the source to target view, and each pixel in the warped image $I_{s\to t}$ can find its matching pixel in the source image using the following equation:

$$p_s = KT_{t\to s}D_t(p_{s\to t})K^{-1}p_{s\to t}, \tag{1}$$

where p_s and $p_{s\to t}$ are the pixel's homogeneous coordinates in I_s and $I_{s\to t}$, respectively, $D(p)$ means the depth value of D at the position p, and K denotes the given camera intrinsic matrix. With the pixel matching relationship, $I_{s\to t}$ can be obtained by filling the color at each pixel position using differentiable bilinear sampling (Jaderberg et al., 2015):

$$I_{s\to t}(p_{s\to t}) = \texttt{BilinearSampler}(I_s(p_s)), \tag{2}$$

where $I(p)$ means the pixel intensity of I at the position p. Since $p_{s\to t}$ is discrete and p_s is continuous, `BilinearSampler` is utilized to calculate the intensity of each pixel in $I_{s\to t}$ using the neighboring pixels around p_s and allow error backpropagation.

After warping, DepthNet and PoseNet can be constrained via the photometric loss, which is calculated as follows:

$$\mathcal{L}_{pht}(I_t, I_{s\to t}) = \alpha \cdot \frac{1-\text{SSIM}(I_t, I_{s\to t})}{2} + (1-\alpha)\cdot|I_t - I_{s\to t}|, \tag{3}$$

where Structure Similarity Index Measure (SSIM) (Wang et al., 2004) constrains the image quality and L1 loss constrains the content. α is set to 0.85 according to Godard et al. (2017).

One of the assumptions for the photometric constraint in Eq. (3) is that the scene contains no specular reflections, and is temporally consistent in terms of brightness (Shao et al., 2022). However, such assumption hardly holds true in the endoscopic cases as visualized in Fig. 1.

3.2. Photometric-invariant Cycle Warping

To overcome the brightness inconsistency, we construct a cycle loop path involving forward and backword warping, as shown in Fig. 2. To better explain the principle, we use the red and blue box contours to indicate the different brightness patterns carried by the image.

To get a warped image inheriting the target's brightness, we start from the target itself, and first warp it to get $I_{t\to s}$, and then warp back to get $I_{t\rightleftarrows s}$. The pixel matching relationships across the three images are formulated as follows:

$$\begin{cases} p_t = KT_{s\to t}D_s(p_{t\to s})K^{-1}p_{t\to s}, \\ p_{t\to s} = KT_{t\to s}D_t(p_{t\rightleftarrows s})K^{-1}p_{t\rightleftarrows s}, \end{cases} \tag{4}$$

where p_t, $p_{t\to s}$, and $p_{t\rightleftarrows s}$ are the pixel's homogeneous coordinates in I_t, $I_{t\to s}$ and $I_{t\rightleftarrows s}$, respectively. D_s and D_t are DepthNet-predicted depth maps of the source and target images, respectively.

Therefore, according to Eq. (2), $I_{t\to s}$ and $I_{t\rightleftarrows s}$ can be obtained sequentially:

$$\begin{cases} I_{t\to s}(p_{t\to s}) = \texttt{BilinearSampler}(I_t(p_t)), \\ I_{t\rightleftarrows s}(p_{t\rightleftarrows s}) = \texttt{BilinearSampler}(I_{t\to s}(p_{t\to s})). \end{cases} \tag{5}$$

Thus, we rewrite the previous photometric constraint Eq. (3) to a cycle form, i.e., $\mathcal{L}_{pht}(I_t, I_{t\rightleftarrows s})$, where I_t and $I_{t\rightleftarrows s}$ have the same brightness pattern as shown in Fig. 2. However, direct usage of such cycle warping brings two issues in the optimization of photometric constraint: (1) noticeable image detail lost due to twice image interpolation using `BilinearSampler`, which negatively affects the appearance-based photometric constraint, and (2) the networks learn too actively to give stable intermediate warping, making the training of two networks difficult to converge. Thus, we introduce two enabling techniques, i.e., structure transplant module to retain structure details and EMA strategy to stabilize forward warping.

3.3. Structure Transplant to Retain Details

The learning-free structure transplant module (STM) works based on two observations: (1) the warped frame $I_{t\to s}$ has homologous image style as I_t, and (2) the original frame I_s has real structure details, and also is roughly aligned with $I_{t\to s}$. Inspired by the image-aligned style transformation proposed in Lv et al. (2023), we transplant fine structure of I_s onto the appearance of $I_{t\to s}$ via a learning-free style transformation approach illustrated in Fig. 3.

Specifically, STM utilizes Fast Fourier Transform (FFT) (Nussbaumer and Nussbaumer, 1982) to decompose an RGB image x into two components, i.e., amplitude $\mathcal{A}(x)$ and phase $\mathcal{P}(x)$. $\mathcal{A}(x)$ mainly controls the image style like brightness and $\mathcal{P}(x)$ contains the information of structure details (Lv et al., 2023). Thus, after the forward warping, STM recombines $\mathcal{A}(I_{t\to s})$ and $\mathcal{P}(I_s)$, and produces the structure-restored warped image $I^{STM}_{t\to s}$ via Inverse Fast Fourier Transform (IFFT):

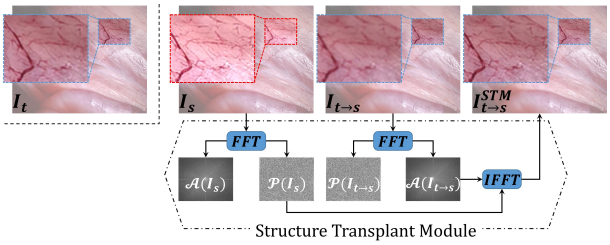

Fig. 3. Details of STM, which utilizes the phase-frequency of the source image I_s to replace that of the warped image $I_{t\to s}$ to avoid image detail lost.

$$I^{STM}_{t\to s} = IFFT(\mathcal{A}(I_{t\to s}) * e^{-j*\mathcal{P}(I_s)}). \tag{6}$$

Therefore, in the backward warping, we utilize $I^{STM}_{t\to s}$ to substitute $I_{t\to s}$ in Eq. (5), which can be rewritten as follows:

$$I_{t\rightleftarrows s}(p_{t\rightleftarrows s}) = \texttt{BilinearSampler}(I^{STM}_{t\to s}(p_{t\to s})). \tag{7}$$

As shown in Fig. 3, $I^{STM}_{t\to s}$ displays better structure details than $I_{t\to s}$, and also maintains similar illumination as $I_{t\to s}$.

3.4. EMA to Stabilize Forward Warping

As shown in Fig. 2, the two paths are connected in a cascaded fashion, where the result of the second warping relies on the output of the previous warping. Therefore, a well-initialized and steady intermediate warped image $I_{t\to s}$ is very important for a stable convergence of training. In view of this, we introduce an exponential moving average (EMA) to bridge the two paths and thus divide the whole training into warm-up and follow-up phases.

During the warm-up training, EMA is disabled and we only train DepthNet and PoseNet in the forward warping by optimizing the photometric constraint between I_s and $I_{t\to s}$, that is, $\theta = \arg\min_\theta \mathcal{L}_{pht}(I_s, I_{t\to s})$, where θ represents the total learnable parameters of both DepthNet and PoseNet. Although the brightness inconsistency occurs in the warm-up training, a reasonably acceptable initialization of DepthNet and PoseNet can be reached.

The follow-up training with EMA enabled is the key step to resist brightness inconsistency based on the cycle warping. We duplicate the network parameters learned in the warm-up training as an EMA copy $\theta_{EMA} \leftarrow \theta$, and use the models with θ_{EMA} to estimate depth and pose in the forward warping and those with θ in the backward warping. θ is updated actively by minimizing the cycle-form photometric loss $L_{pht}(I_t, I_{t\rightleftarrows s})$, and θ_{EMA} is updated slowly by moving averaging θ. That is, the copy of DepthNet and PoseNet for the forward warping is **not** directly learned by the optimization of photometric constraint in the follow-up training, thus to avoid predicting frequently changed $I_{t\to s}$.

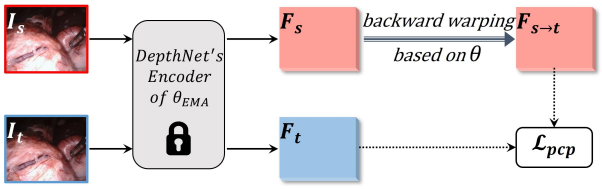

Fig. 4. The auxiliary perception constraint by backward warping the encoding feature maps instead of raw images.

Besides, we further enhance the learning robustness via a ready-made perception loss. As shown in Fig. 4, we backward warp the encoding map of I_s generated from DepthNet's encoder of its EMA version, yielding a warped feature map $F_{s\to t}$. If we also extract the encoding map of I_t using the same encoder, $F_{s\to t}$ and F_t should be aligned and may contain high-level semantic information covering the low-level brightness changes. Thus, a perception loss is defined:

$$\mathcal{L}_{pcp}(F_t, F_{s\to t}) = |F_t - F_{s\to t}|. \tag{8}$$

In summary, the follow-up training of MonoPCC is formulated as follows:

$$\begin{cases} \theta = \arg\min_{\theta}(\mathcal{L}_{pht}(I_t, I_{t\rightleftarrows s}) + \mathcal{L}_{pcp}(F_t, F_{s\to t})), \\ \theta_{EMA} = \alpha\theta_{EMA} + (1-\alpha)\theta, \end{cases} \tag{9}$$

where α is a hyperparameter to balance the stability and learnability of the forward warping, and set to 0.75 in MonoPCC by default.

3.5. Implementation Details

We train MonoPCC using a single NVIDIA RTX A6000 GPU. We utilize MonoViT and ResNet18 pre-trained on ImageNet(Deng et al., 2009) for DepthNet and PoseNet initialization, respectively. We set the batch size to 12, and use AdamW (Loshchilov and Hutter, 2017) as the optimizer. The learning rate decay is exponential and set to 0.9. The number of epochs and learning rate differ in the warm-up and follow-up training phases. In warm-up training, the number of epochs is set to 20 epochs, and the initial learning rate is set to 1×10^{-4} for DepthNet's encoder and 5×10^{-5} for the rest weights. In follow-up training, the number of epochs is set to 10 epochs, and the initial learning rate is set to 5×10^{-5} for all network weights.

4. Experimental Settings

4.1. Comparison Methods

We compare MonoPCC with 8 state-of-the-art (SOTA) methods of monocular depth estimation, i.e., Monodepth2 (Godard et al., 2019), FeatDepth (Shu et al., 2020), HR-Depth (Lyu et al., 2021), DIFFNet (Zhou et al., 2021), Endo-SfMLearner (Ozyoruk et al., 2021), AF-SfMLearner (Shao et al., 2022), MonoViT (Zhao et al., 2022), and Lite-Mono (Zhang et al., 2023). We evaluate their performance and make comparisons using their released codes.

Table 1. Evaluation metrics of monocular depth estimation, where N refers to the number of valid pixels in depth maps, d_i and d_i^* denote the estimated and GT depth of i-th pixel, respectively. The Iverson bracket $[\cdot]$ yields 1 if the statement is true, otherwise 0.

Metric	Formula
Abs Rel	$\frac{1}{N}\sum_{i=0}^{N-1} \lvert d_i - d_i^* \rvert / d_i^*$
Sq Rel	$\frac{1}{N}\sum_{i=0}^{N-1} \lvert d_i - d_i^* \rvert^2 / d_i^*$
RMSE	$\sqrt{\frac{1}{N}\sum_{i=0}^{N-1} \lvert d_i - d_i^* \rvert^2}$
RMSE log	$\sqrt{\frac{1}{N}\sum_{i=0}^{N-1} \lvert \log d_i - \log d_i^* \rvert^2}$
δ	$\frac{1}{N}\sum_{i=0}^{N-1} [max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) < 1.25]$

4.2. Datasets

The experiment in this work involves 4 public endoscopic datasets, i.e., SCARED (Allan et al., 2021), SimCol3D (Rau et al., 2023), SERV-CT (Edwards et al., 2022), and Hamlyn (Mountney et al., 2010; Stoyanov et al., 2010; Pratt et al., 2010), and a public outdoor-scene dataset, i.e., KITTI (Geiger et al., 2012).

SCARED is from a MICCAI challenge and consists of 35 videos with the size of 1280×1024 collected from porcine cadavers using a da Vinci Xi surgical robot. Structured light is used to obtain the ground-truth (GT) depth map for the first frame, which is then mapped onto the subsequent frames using camera trajectory recorded by the robotic arm.

Since the original data is binocular, we only use the left view to simulate our focusing monocular situation. We split the dataset into 21,066 frames for training and 551 for test, which are consistent with the previous SOTA (Shao et al., 2022). No videos are cross-used in both training and test.

SimCol3D is from MICCAI 2022 EndoVis challenge. It is a synthetic dataset which contains over 36,000 colonoscopic images and depth annotations with size of 475×475. Meanwhile, virtual light sources are attached to the camera. For the usage of SimCol3D dataset, we follow their official website and spilt the dataset into 28,776 and 9,009 frames for training and test, respectively.

SERV-CT is collected from two *ex vivo* porcine cadavers, and each has 8 binocular keyframes. We treat the two views independently and thus have 32 images in total with the size of 720×576. The GT depth map for each image is calculated by manually aligning the endoscopic image to the 3D anatomical model derived from the corresponding CT scan.

Hamlyn is a large public endoscopic dataset with multiple stereo videos for clinical application, including roughly 8,000 frames with different resolutions. Recasens et al. (2021) provided a rectified version and created GT depth maps for left-view frames using software *Libelas*.

KITTI is an outdoor dataset captured in autonomous driving scenarios, e.g., road and campus, while laser sensors are

Table 2. Quantitative comparison results on SCARED and SimCol3D. The best results are marked in bold and the second-best underlined. The paired *p*-values between MonoPCC and others are all less than 0.05.

Methods	SCARED					SimCol3D				
	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑
Monodepth2	0.060	0.432	4.885	0.082	0.972	0.076	0.061	0.402	0.106	0.950
FeatDepth	0.055	0.392	4.702	0.077	0.976	0.077	0.069	0.374	0.098	0.957
HR-Depth	0.058	0.439	4.886	0.081	0.969	0.072	0.044	0.378	0.100	0.961
DIFFNet	0.057	0.423	4.812	0.079	0.975	0.074	0.053	0.401	0.105	0.957
Endo-SfMLearner	0.057	0.414	4.756	0.078	0.976	0.072	0.042	0.407	0.103	0.950
AF-SfMLearner	0.055	0.384	4.585	0.075	0.979	0.071	0.045	0.372	0.099	0.961
MonoViT	0.057	0.416	4.919	0.079	0.977	0.064	0.034	0.377	0.094	0.968
Lite-Mono	0.056	0.398	4.614	0.077	0.974	0.076	0.050	0.424	0.110	0.950
MonoPCC(Ours)	**0.051**	**0.349**	**4.488**	**0.072**	**0.983**	**0.058**	**0.028**	**0.347**	**0.090**	**0.975**

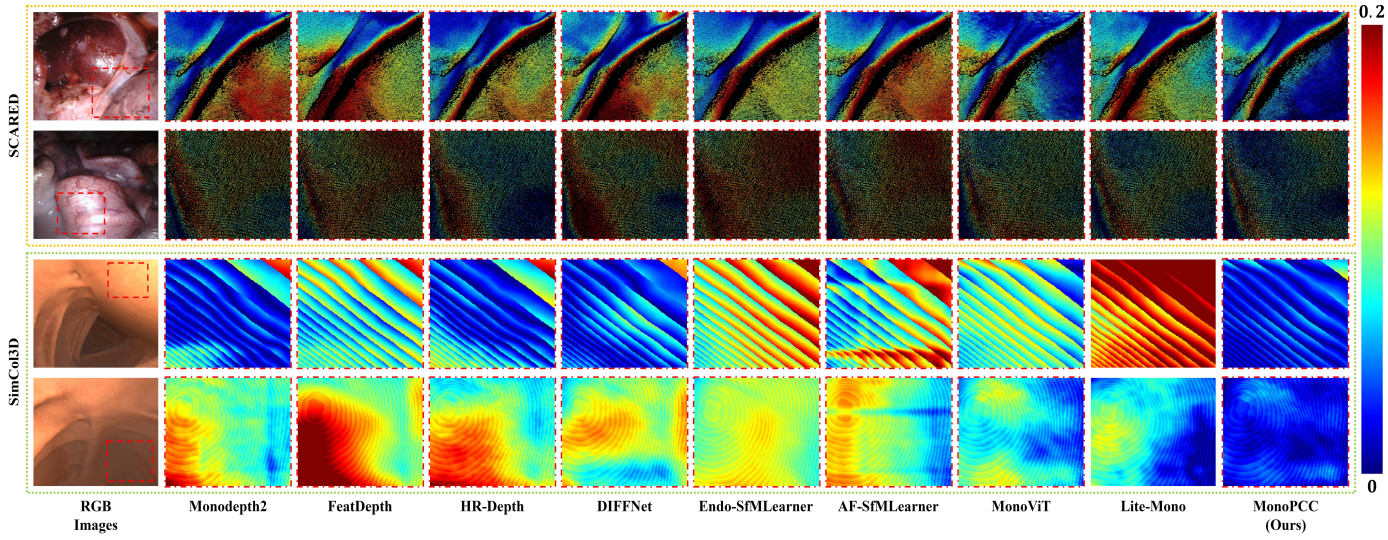

Fig. 5. The Abs Rel error maps of comparison methods on SCARED and SimCol3D. Opencv Jet Colormap is used for visualization.

equipped to scan precise depth maps. We follow the split of Eigen et al. (2014) and Zhou et al. (2017), and use 39,810 frames for training, 4,424 for validation, and 697 for test.

4.3. Evaluation Metrics

To be consistent with previous works, we employ 5 metrics in the evaluation, which are listed in Table 1.

Note that, as the common evaluation routine of monocular depth estimation, the predicted depth map should be scaled beforehand since it is scale unknown. The scaling factor is the GT median depth divided by the predicted median depth (Zhou et al., 2017).

Also, for a fair comparison, we follow the previous works (Shao et al., 2022; Rau et al., 2023; Recasens et al., 2021; Godard et al., 2019), and cap the depth maps within the maximum value for each dataset. Specifically, the maximum depth value is set to 150 *mm*, 200 *mm*, 180 *mm*, 300 *mm*, and 80 *m* for SCARED, SimCol3D, SERV-CT, Hamlyn and KITTI, respectively.

Furthermore, for each comparison in the following experimental results, we perform paired samples T-test (Duncan, 1975) and report the *p*-value to indicate whether the difference is significant.

5. Results and Discussions

5.1. Comparison with State-of-the-arts

5.1.1. Evaluation on SCARED and SimCol3D

Table 2 shows the comparison results on SCARED (left part) and SimCol3D (right part). As can be seen, our method achieves the best performance in terms of all five metrics on SCARED, and consistently outperforms the other SOTAs by different margins on SimCol3D.

Likewise, Endo-SfMLearner and AF-SfMLearner also belong to the approach of addressing the brightness inconsistency in self-supervised monocular depth estimation. Endo-SfMLearner adopts the simple idea of brightness linear transformation, while AF-SfMLearner resorts to a more complex appearance flow model. From their results, we can see that the appearance flow model is more effective, since the brightness changes are often non-linear in the endoscopic scene. Nevertheless, our MonoPCC still exceeds the AF-SfMLearner by

Table 3. Quantitative comparison results on SERV-CT and Hamlyn. The best results are marked in bold and the second-best underlined. The paired *p*-values between MonoPCC and others are all less than 0.05, except for δ on SERV-CT compared to MonoViT.

Methods	SERV-CT					Hamlyn				
	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑
Monodepth2	0.127	2.152	13.023	0.166	0.825	0.087	1.254	9.555	0.115	0.939
FeatDepth	0.117	1.862	12.040	0.154	0.841	0.064	0.939	8.016	0.090	0.964
HR-Depth	0.122	2.085	12.587	0.156	0.850	0.076	1.139	8.883	0.102	0.957
DIFFNet	0.116	1.858	12.177	0.146	0.864	0.070	1.093	8.552	0.097	0.958
Endo-SfMLearner	0.122	2.123	12.551	0.168	0.842	0.063	0.886	7.901	0.091	0.968
AF-SfMLearner	0.101	1.546	10.900	0.131	0.888	0.078	1.018	8.712	0.104	0.968
MonoViT	0.103	1.566	11.482	0.136	0.895	0.073	0.945	8.470	0.099	0.975
Lite-Mono	0.124	2.314	13.156	0.175	0.820	0.074	1.106	8.743	0.104	0.950
MonoPCC(Ours)	**0.091**	**1.252**	**10.059**	**0.116**	**0.915**	**0.061**	**0.819**	**7.604**	**0.085**	**0.977**

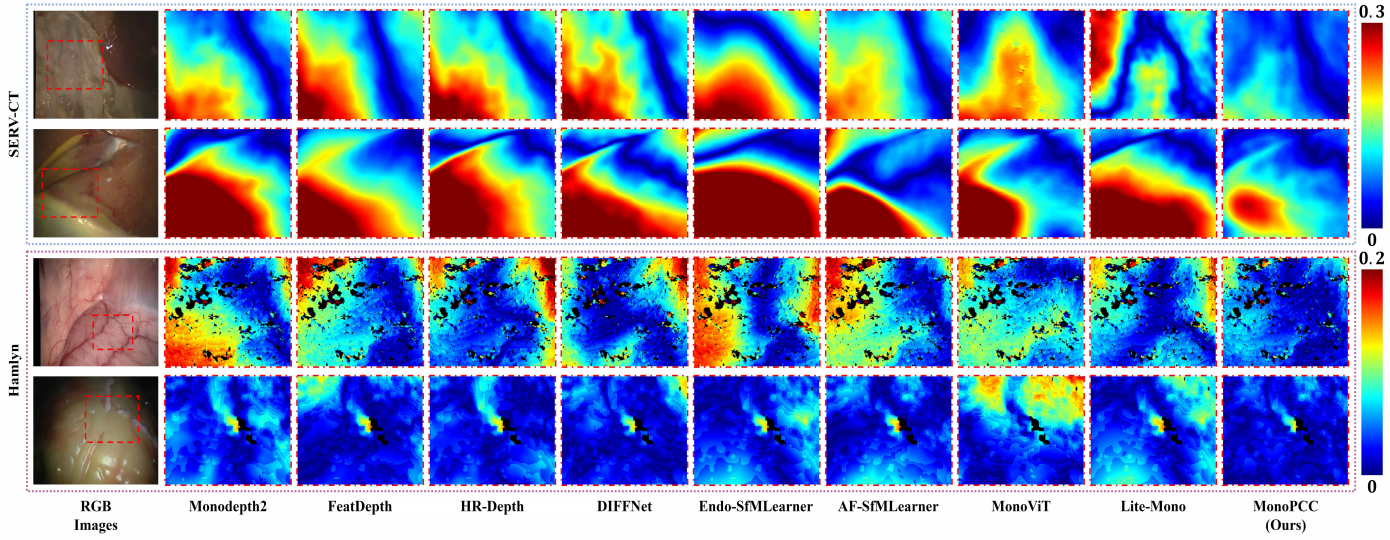

Fig. 6. The Abs Rel error maps of comparison methods on SERV-CT and Hamlyn. Opencv Jet Colormap is used for visualization.

reducing Sq Rel by 9.11% on SCARED, which implies that compared to the learning-based appearance flow model, our cycle-form warping can guarantee the brightness consistency in a more effective and reliable way.

Fig. 5 presents a qualitative comparison on the two datasets. Error maps are acquired by mapping pixel-wise Abs Rel value to different colors, and the red means large relative error and the blue indicates small error. On SCARED, the color-coded error map of MonoPCC is clearly bluer than those of other SOTAs, especially for the regions near the boundaries or specular reflections, as indicated by the red dotted boxes in Fig. 5. Specifically, the failure of other methods in these cases can be attributed to neglect or misguidance on brightness inconsistency, leading DepthNet to learn inappropriate depth prior knowledge. On SimCol3D, MonoPCC provides lower relative depth errors at both the proximal and distal parts of the colon, which demonstrates the validity of MonoPCC for low-textured regions, e.g., colon in the digestive tract.

5.1.2. Generalization on SERV-CT and Hamlyn

We also evaluate the generalization of all methods, i.e., using the model trained on SCARED to directly test on SERV-CT and Hamlyn. The comparison results are listed in Table 3.

As can be seen, all methods experience a varying degree of degradation on both datasets. However, on SERV-CT (left part of Table 3), MonoPCC is still the best, and the only one maintaining the Abs Rel less than 0.1 and δ greater than 0.9. Also, MonoPCC significantly surpasses the second-best AF-SfMLearner by a relative decrease of 19.02% and 11.45% in terms of the Sq Rel and RMSE log ($p < 0.005$), respectively.

Compared to the results on SERV-CT, the performance degradation on Hamlyn is somewhat not that severe. We believe that this is because the captured scenes of Hamlyn are relatively homogeneous, without complex and irregular surfaces as SERV-CT. Compared to the second-best Endo-SfMLearner, MonoPCC achieves the lowest error in terms of the first four metrics, and the highest percentage of the inliers in terms of δ ($p < 0.001$ for all metrics).

Fig. 6 exhibits four visual results from SERV-CT and Hamlyn, respectively. As can be seen, compared to the other methods, MonoPCC produces less red error maps as indicated by the dotted boxes in Fig. 6. The two cases from Hamlyn contain motion blur during *in vivo* porcine procedure (the 3^{rd} row),

Table 4. The rows except the first one are the comparison results of the five variants and the complete MonoPCC, which are all cycle-constrained. The first row is the backbone MonoViT using the regular non-cycle constraint.

Cycle	STM	EMA	$\mathcal{L}_{pcp}$	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
○	✗	✗	✗	0.106±0.051	1.533±1.383	9.737±5.221	0.141±0.060
●	✗	✗	✗	0.110±0.047	1.584±1.315	10.003±4.930	0.145±0.056
●	✗	✓	✓	0.108±0.048	1.596±1.329	10.047±5.335	0.144±0.060
●	✓	✗	✗	0.099±0.043	1.286±1.091	9.007±4.697	0.131±0.052
●	✓	✗	✓	0.091±0.031	1.040±0.700	8.239±3.664	0.120±0.041
●	✓	✓	✗	0.091±0.036	1.081±0.800	8.355±3.905	0.121±0.045
●	✓	✓	✓	**0.085±0.034**	**0.953±0.712**	**7.734±3.537**	**0.114±0.044**

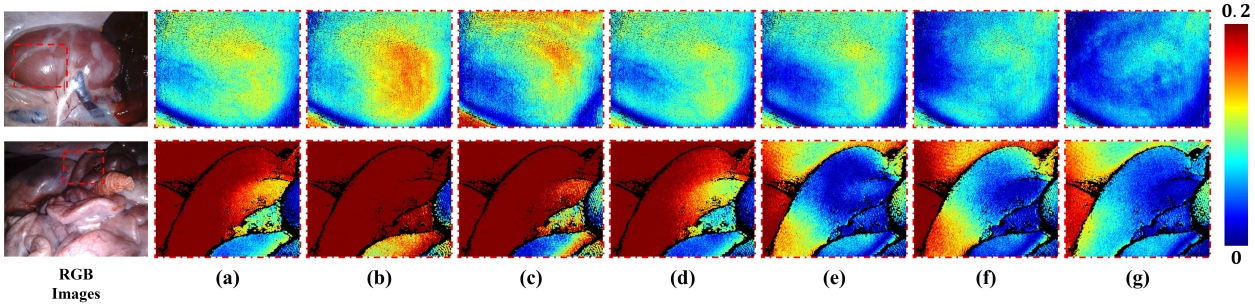

Fig. 7. The Abs Rel error maps of seven ablation variants, including effectiveness of three components. (a)-(g) correspond to the 1^{st} to the 7^{th} rows in Table 4. Opencv Jet Colormap is used for visualization.

and brightness fluctuations in the over-smoothed areas of silicon heart (the 4^{th} row). The color-coded error maps on the two challenging cases verify that MonoPCC displays almost deep blue error maps in the highlighted regions, and has better generalization on the non-rigid and reflective tissues.

5.2. Ablation Study

We conduct ablation studies via 5-fold cross-validation on SCARED to verify the effectiveness of each component in MonoPCC and to compare MonoPCC with other related techniques against brightness fluctuations.

5.2.1. Effectiveness of Three Components

We develop five variants of MonoPCC by disabling the three components, EMA and/or $\mathcal{L}_{pcp}$, and STM. Note that, the warping should be cycle-form to utilize the components, and if we disable EMA, only θ in the backward warping path will be updated, and θ_{EMA} in the forward warping remains unchanged. Table 4 and Fig. 7 give the comparison results between the variants and complete MonoPCC. We also include the backbone MonoViT in the first row of Table 4 as the baseline, which adopts the non-cycle warping.

From the comparison results in Table 4 and visualization cases in Fig. 7, three key observations can be made:

(1) By comparing the first two rows of Table 4 and Fig. 7 (a)-(b), the inferior performance of the variant using none of the components indicates that the direct usage of cycle warping is not enough due to the image blurring and unstable gradient propagation mentioned in Sec. 3.2.

(2) As can be seen from the second to fourth rows of Table 4 and Fig. 7 (b)-(d), using EMA and $\mathcal{L}_{pcp}$ solely brings no improvements, which still creates much unsatisfactory estimation in the red dotted boxes of Fig. 7 (c). However, adding STM can significantly reduce Sq Rel and RMSE by 16.11% and 7.50%, respectively ($p < 0.003$). Thus, STM is the irreplaceable component for successfully training using the cycle-form warping.

(3) Comparison between the last four rows of Table 4 and Fig. 7 (d)-(g) reveals the effectiveness of EMA and $\mathcal{L}_{pcp}$. Specifically, with STM enabled, both EMA and $\mathcal{L}_{pcp}$ bring remarkable reduction of Abs Rel by 8.08% ($p = 0.012$, $p = 0.022$). Meanwhile, EMA and $\mathcal{L}_{pcp}$ are not mutually excluded, and the combination reduces Abs Rel by 14.14% ($p = 0.001$). As illustrated in Fig. 7 (g), there exist more blue-tuned pixels in the red dotted regions of interest (i.e., either light or dark areas).

5.2.2. MonoPCC vs. Other Techniques against Brightness Fluctuations

Several techniques have been developed to address inconsistent brightness in endoscopic images, e.g., affine brightness transformation (ABT) in Endo-SfMLearner, and appearance flow module (AFM) in AF-SfMLearner. For comparison, we train three variants of the backbone MonoViT, using ABT, AFM, and STM+EMA, respectively, to verify the robustness of different enabling techniques to the brightness inconsistency.

Table 5 lists the comparison results as well as the backbone MonoViT using none of these techniques. As can be seen, the variant using STM+EMA strategy exceeds that using AFM only with a reduction of Abs Rel and Sq Rel by 7.14% and 13.59%, respectively ($p = 0.019$, $p = 0.014$). Fig. 8 displays two visualization examples in challenging cases, e.g., the boundary and flat area of tissues. As shown in Fig. 8, all techniques for addressing the brightness fluctuation can help to gain varying degrees of improvements. Furthermore, we observe that our strategy performs better than both ABT and AFM. It is worth

Table 5. Comparison results of different techniques for addressing the brightness fluctuations in self-supervised learning. The last row is the technique used in MonoPCC.

Schemes	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
Baseline	0.106±0.051	1.533±1.383	9.737±5.221	0.141±0.060
ABT	0.102±0.046	1.425±1.214	9.366±4.784	0.137±0.056
AFM	0.098±0.039	1.251±0.964	8.897±4.203	0.130±0.049
STM+EMA	**0.091±0.036**	**1.081±0.800**	**8.355±3.905**	**0.121±0.045**

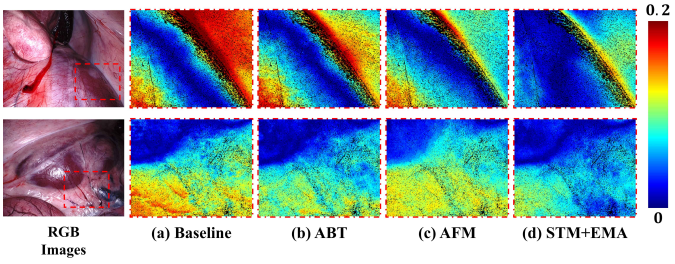

Fig. 8. The Abs Rel error maps of MonoPCC and other similar modules against photometric inconsistency. (a)-(d) correspond to the 1^{st} to the 4^{th} rows in Table 5. Opencv Jet Colormap is used for visualization.

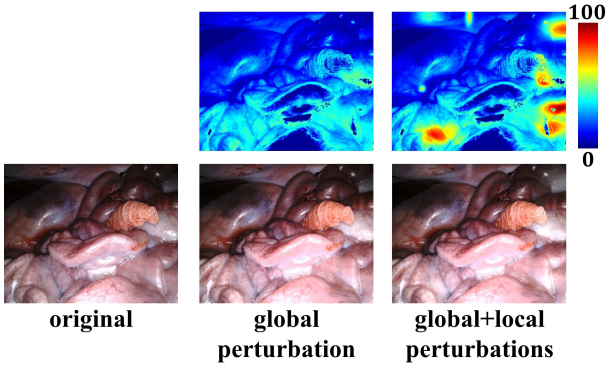

Fig. 9. An example of created brightness perturbation. From left to right is the original image, globally perturbated ($k = 1.2$), and both globally and locally (bright spots) perturbated. The color-coded maps above them describe the subtractive difference between the perturbated image and its original one.

noting that compared to AFM, our method requires no extra model to learn, exhibiting greater usability and scalability.

5.3. Robustness to Severe Brightness Inconsistency

Besides the current inconsistent brightness carried by the used dataset, we want to test the limit of MonoPCC under more severe brightness fluctuations. To this end, we create two copies of SCARED, and add global brightness perturbation to every adjacent frames of the first copy, and both global and local perturbations to the second. Specifically, the global perturbation is defined as the linear transformation on the image's brightness channel (HSV color model), that is, $v^{scale} = k * v, k \in [0.8, 0.9] \cup [1.1, 1.2]$. The local perturbation is the randomly placed Gaussian bright (or dark) spots. Fig. 9 presents visual examples from the original SCARED and our created two copies with more severe brightness inconsistency.

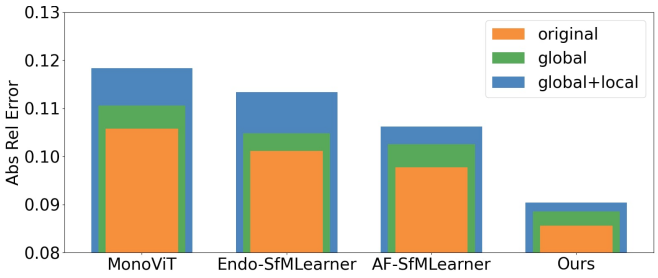

Fig. 10. The Abs Rel errors of different methods trained on the two brightness-perturbed copies of SCARED and the original SCARED.

Fig. 10 illustrates Abs Rel errors of different methods on the two brightness-perturbed copies of and the original SCARED. As mentioned before, Endo-SfMLearner and AF-SfMLearner all tried to address the brightness inconsistency. As can be seen from the comparison results, Endo-SfMLearner shows a robustness to the global brightness perturbation thanks to its using brightness linear transformation, but degrades significantly when facing the local perturbation. AF-SfMLearner can address the local brightness perturbation more or less using a learned appearance flow model, but still presents a non-trivial increase of Abs Rel error. In comparison, MonoPCC achieves the lowest Abs Rel error on the three datasets, and even produces lower errors on the copy containing the most severe brightness inconsistency (global+local) compared to the other methods' results on the original SCARED.

5.4. Results on KITTI Benchmark

Besides the demonstration on the medical scenarios, i.e., endoscopic images, we in this subsection verify the competitiveness of MonoPCC on the popular natural-scence benchmark, i.e., KITTI (Geiger et al., 2012). KITTI is an outdoor dataset, which roughly obeys brightness consistency assumption due to constant illumination condition, i.e., the sunshine. However, some factors like specular reflection on the car body also create some negative impacts on the ideal photometric supervision.

Table 6 provides the comparison results with the previously mentioned SOTAs. Most of them have reported their evaluation results on KITTI, so we directly copy their results. For the two methods, i.e., Endo-SfMLearner and AF-SfMLearner, which have no official results on KITTI, we use their released codes and default settings to train two models on KITTI.

As indicated in Table 6, MonoPCC and MonoViT are the only two methods whose Abs Rel values are less than 0.1. Moreover, MonoPCC outperforms MonoViT in terms of all

Table 6. Quantitative comparison results on KITTI. The best results are marked in bold and the second best ones are underlined.

	Year	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑
Monodepth2	2019	0.115	0.903	4.863	0.193	0.877
FeatDepth	2020	0.104	0.729	4.481	0.179	0.893
HR-Depth	2021	0.109	0.792	4.632	0.185	0.884
DIFFNet	2021	0.102	0.749	4.445	0.179	0.897
Endo-SfMLearner	2021	0.100	0.722	4.457	0.178	<u>0.898</u>
AF-SfMLearner	2022	0.113	1.110	4.949	0.189	0.883
MonoViT	2022	<u>0.099</u>	<u>0.708</u>	<u>4.372</u>	<u>0.175</u>	**0.900**
Lite-Mono	2023	0.101	0.729	4.454	0.178	0.897
MonoPCC(Ours)		**0.098**	**0.677**	**4.318**	**0.173**	**0.900**

metrics, especially with a relative decrease of 4.38% in terms of Sq Rel error ($p < 0.001$).

The two endoscopy-tailored methods, i.e., Endo-SfMLearner and AF-SfMLearner, are the second or the third best on the four endoscopic datasets, but unexpectedly slip several places on KITTI. We believe the potential reason is that the two methods are overly focusing on the brightness adjustment, and try to change the image appearance no matter the brightness is inconsistent or not. Since KITTI has no that significant brightness fluctuations, such forcible brightness adjustment could induce errors more or less. In comparison, our MonoPCC is essentially making the brightness patterns inherit across the source and target images, and thus the strength of brightness adjustment can be naturally compatible with the true situation of brightness fluctuations. Therefore, MonoPCC shows the consistent competitiveness in both medical and natural scenes.

6. Conclusion

Self-supervised monocular depth estimation is challenging for endoscopic scenes due to the severe negative impact of brightness fluctuations on the photometric constraint. In this paper, we propose a cycle-form warping to naturally overcome the brightness inconsistency of endoscopic images, and develop a MonoPCC for robust monocular depth estimation by using a re-designed photometric-invariant cycle constraint. To make the cycle-form warping effective in the photometric constraint, MonoPCC is equipped with two enabling techniques, i.e., structure transplant module (STM) and exponential moving average (EMA) strategy. STM alleviates image detail degradation to validate the backward warping, which uses the result of forward warping as input. EMA bridges the learning of network weights in the forward and backward warping, and stabilizes the intermediate warped image to ensure an effective convergence. The comprehensive and extensive comparisons with 8 state-of-the-arts on five public datasets, i.e., SCARED, SimCol3D, SERV-CT, Hamlyn, and KITTI, demonstrate that MonoPCC achieves a superior performance by decreasing the absolute relative error by at least 7.27%, 9.38%, 9.90% and 3.17% on the four endoscopic datasets, respectively, and shows the competitiveness even for the natural scenario. Additionally, two ablation studies are conducted to confirm the effectiveness of three developed modules and the advancement of MonoPCC over other similar techniques against brightness fluctuations. However, the current pipeline only leverages a single frame to infer the depth map, which is likely to lack context information and geometric consistency. In the future work, we plan to extend MonoPCC into the multi-frame based monocular depth estimation.

Acknowledgments

This work was supported in part by National Key R&D Program of China (Grant No. 2023YFC2414900), Fundamental Research Funds for the Central Universities (2021XXJS033), Research grants from United Imaging Healthcare Inc.

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 .

Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I., 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in neural information processing systems 32.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Duncan, D.B., 1975. T tests and intervals for comparisons suggested by the data. Biometrics , 339–359.

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

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems 27.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE. pp. 3354–3361.

Godard, C., Mac Aodha, O., Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 270–279.

Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J., 2019. Digging into self-supervised monocular depth estimation, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3828–3838.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, B., Zheng, J.Q., Nguyen, A., Xu, C., Gkouzionis, I., Vyas, K., Tuch, D., Giannarou, S., Elson, D.S., 2022. Self-supervised depth estimation in laparoscopic image using 3d geometric consistency, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 13–22.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. Advances in neural information processing systems 28.

Li, W., Hayashi, Y., Oda, M., Kitasaka, T., Misawa, K., Mori, K., 2022. Geometric constraints for self-supervised monocular depth estimation on laparoscopic images with dual-task consistency, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 467–477.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 .

Lv, J., Zeng, X., Wang, S., Duan, R., Wang, Z., Li, Q., 2023. Robust one-shot segmentation of brain tissues via image-aligned style transformation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1861–1869.

Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y., 2021. Hr-depth: High resolution self-supervised monocular depth estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2294–2301.

Mountney, P., Stoyanov, D., Yang, G.Z., 2010. Three-dimensional tissue deformation recovery and tracking. IEEE Signal Processing Magazine 27, 14–24.

Nussbaumer, H.J., Nussbaumer, H.J., 1982. The fast Fourier transform. Springer.

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.

Pratt, P., Stoyanov, D., Visentini-Scarzanella, M., Yang, G.Z., 2010. Dynamic guidance for robotic surgery using image-constrained biomechanical models, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010: 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part I 13, Springer. pp. 77–85.

Rau, A., Bhattarai, B., Agapito, L., Stoyanov, D., 2023. Bimodal camera pose prediction for endoscopy. IEEE Transactions on Medical Robotics and Bionics .

Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J., 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, 7225–7232.

Ruhkamp, P., Gao, D., Chen, H., Navab, N., Busam, B., 2021. Attention meets geometry: Geometry guided spatial-temporal attention for consistent self-supervised monocular depth estimation, in: 2021 International Conference on 3D Vision (3DV), IEEE. pp. 837–847.

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

Shu, C., Yu, K., Duan, Z., Yang, K., 2020. Feature-metric loss for self-supervised learning of depth and egomotion, in: European Conference on Computer Vision, Springer. pp. 572–588.

Stoyanov, D., Scarzanella, M.V., Pratt, P., Yang, G.Z., 2010. Real-time stereo reconstruction in robotically assisted minimally invasive surgery, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010: 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part I 13, Springer. pp. 275–282.

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

Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M., 2021. The temporal opportunist: Self-supervised multi-frame monocular depth, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1164–1174.

Yin, Z., Shi, J., 2018. Geonet: Unsupervised learning of dense depth, optical flow and camera pose, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1983–1992.

Zhang, N., Nex, F., Vosselman, G., Kerle, N., 2023. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18537–18546.

Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S., 2022. Monovit: Self-supervised monocular depth estimation with a vision transformer, in: 2022 International Conference on 3D Vision (3DV), IEEE. pp. 668–678.

Zhou, H., Greenwood, D., Taylor, S., 2021. Self-supervised monocular depth estimation with internal feature fusion. arXiv preprint arXiv:2110.09482 .

Zhou, T., Brown, M., Snavely, N., 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, pp. 1851–1858.