Multi-view Guidance for Self-supervised Monocular Depth Estimation on Laparoscopic Images via Spatio-Temporal Correspondence

Wenda Li[1(✉)], Yuichiro Hayashi[1], Masahiro Oda[1,2], Takayuki Kitasaka[3], Kazunari Misawa[4], and Kensaku Mori[1,5,6(✉)]

[1] Graduate School of Informatics, Nagoya University, Aichi, Nagoya 464-8601, Japan
wdli@mori.m.is.nagoya-u.ac.jp

[2] Information and Communications, Nagoya University, Aichi, Nagoya 464-8601, Japan

[3] Faculty of Information Science, Aichi Institute of Technology, Yakusacho, Aichi, Toyota 470-0392, Japan

[4] Aichi Cancer Center Hospital, Aichi, Nagoya 464-8681, Japan

[5] Information Technology Center, Nagoya University, Aichi, Nagoya 464-8601, Japan

[6] Research Center of Medical Bigdata, National Institute of Informatics, Tokyo, Hitotsubashi 101-8430, Japan
kensaku@is.nagoya-u.ac.jp

Abstract. This work proposes an innovative self-supervised approach to monocular depth estimation in laparoscopic scenarios. Previous methods independently predicted depth maps ignoring spatial coherence in local regions and temporal correlation between adjacent images. The proposed approach leverages spatio-temporal coherence to address the challenges of textureless areas and homogeneous colors in such scenes. This approach utilizes a multi-view depth estimation model to guide monocular depth estimation when predicting depth maps. Moreover, the minimum reprojection error is extended to construct a cost volume for the multi-view model using adjacent images. Additionally, a 3D consistency of the point cloud back-projected from predicted depth maps is optimized for the monocular depth estimation model. To benefit from spatial coherence, deformable patch-matching is introduced to the monocular and multi-view models to smooth depth maps in local regions. Finally, a cycled prediction learning for view synthesis and relative poses is designed to exploit the temporal correlation between adjacent images fully. Experimental results show that the proposed method outperforms existing methods in both qualitative and quantitative evaluations. Our code is available at https://github.com/MoriLabNU/MGMDepthL.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_41.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 429–439, 2023.
https://doi.org/10.1007/978-3-031-43996-4_41

Keywords: Monocular depth estimation · Self-supervised learning · Laparoscopic images

1 Introduction

The significance of depth information is undeniable in computer-assisted surgical systems [20]. In robotic-assisted surgery, the depth value is used to accurately map the surgical field and track the movement of surgical instruments [7,22]. Additionally, the depth value is essential to virtual and augmented reality to create 3D models and realize surgical technique training [18].

Learning-based approaches have significantly improved monocular depth estimation (MDE) in recent years. As the pioneering work, Eigen et al. [2] proposed the first end-to-end deep learning framework using multi-scale CNN under supervised learning for MDE. Following this work, ResNet-based and Hourglass-based models were introduced as the variants of CNN-based methods [9,21]. However, supervised learning requires large amounts of annotated data. Collecting depth value from hardware or synthesis scenes is time-consuming and expensive [14]. To address this issue, researchers have explored self-supervised methods for MDE. Zhou et al. [30] proposed a novel depth-pose self-supervised monocular depth estimation from a video sequence. They generate synthesis views through the estimated depth maps and relative poses. Gordon et al. [3] optimized this work by introducing a minimum reprojection error between adjacent images and made a notable baseline named Monodepth2. Subsequently, researchers have proposed a variety of self-supervised methods for MDE, including those based on semantic segmentation [4], adversarial learning [28] and uncertainty [17]. These days, MDE has been applied to laparoscopic images. Ye et al. [27] and Max et al. [1] have provided exceptional laparoscopic scene datasets. Huang et al. [6] used generative adversarial networks to derive depth maps on laparoscopic images. Li et al. [12] combined depth estimation with scene coordinate prediction to improve network performance.

This study presents a novel approach to predict depth values in laparoscopic images using spatio-temporal correspondence. Current self-supervised models for monocular depth estimation face two significant challenges in laparoscopic settings. First, monocular models individually predicted depth maps, ignoring the temporal correlation between adjacent images. Second, accurate point matching is difficult to achieve due to the misleading of large textureless regions caused by the smooth surface of organs. And the homogenous color misleads that the local areas of the edge are regarded with the same depth value. To overcome these obstacles, We introduce multi-view depth estimation (MVDE) with the optimized cost volume to guide the self-supervised monocular depth estimation model. Moreover, we exploit more informative values in a spatio-temporal manner to address the limitation of existing multi-view and monocular models.

Our main contributions are summarized as follows. (i) A novel self-supervised monocular depth estimation guided by a multi-view depth model to leverage adjacent images when estimating depth value. (ii) Cost volume construction for multi-view depth estimation under minimum reprojection error and an optimized point cloud consistency for the monocular depth estimation. (iii) An extended deformable patch matching based on the spatial coherence in local regions and a cycled prediction learning for view synthesis and relative poses to exploit the temporal correlation between adjacent images.

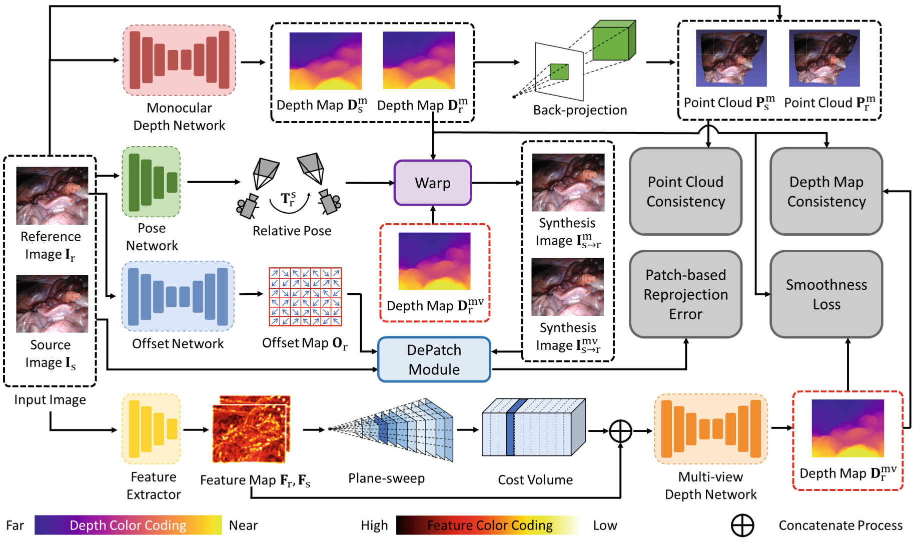

Fig. 1. Overview of our self-supervised monocular depth estimation framework. The proposed method consists of a monocular depth network, a pose network, an offset network, and a multi-view depth network.

2 Method

2.1 Self-supervised Monocular Depth Estimation

Following [3,12], we train a self-supervised monocular depth network using a short clip from a video sequence. The short clip consists of a current frame $\mathbf{I}_t$ as reference image $\mathbf{I}_r$ and adjacent images $\mathbf{I}_s \in \{\mathbf{I}_{t-1}, \mathbf{I}_{t+1}\}$ regarded as source images. As shown in Fig. 1, a monocular depth network $f_D^m(\mathbf{I}_r, \mathbf{I}_s; \theta_D^m)$ respectively predicts pixel-level depth maps $\mathbf{D}_r$ and $\mathbf{D}_s$ corresponding to $\mathbf{I}_r$ and $\mathbf{I}_s$. A pose network $f_T(\mathbf{I}_r, \mathbf{I}_s; \theta_T)$ estimates a transformation matrix $\mathbf{T}_r^s$ as a relative pose of the laparoscope from view $\mathbf{I}_r$ to $\mathbf{I}_s$. We use $\mathbf{D}_r$ and $\mathbf{T}_r^s$ to match the pixels between $\mathbf{I}_r$ and $\mathbf{I}_s$ by

$$\boldsymbol{p}_s = \mathbf{K}\mathbf{T}_r^s\mathbf{D}(\boldsymbol{p}_r)\mathbf{K}^{-1}\boldsymbol{p}_r, \tag{1}$$

where $\boldsymbol{p}_r$ and $\boldsymbol{p}_s$ are the 2D pixel coordinates in $\mathbf{I}_r$ and $\mathbf{I}_s$. $\mathbf{K}$ is the laparoscope's intrinsic parameter matrix. This allows for the generation of a synthetic

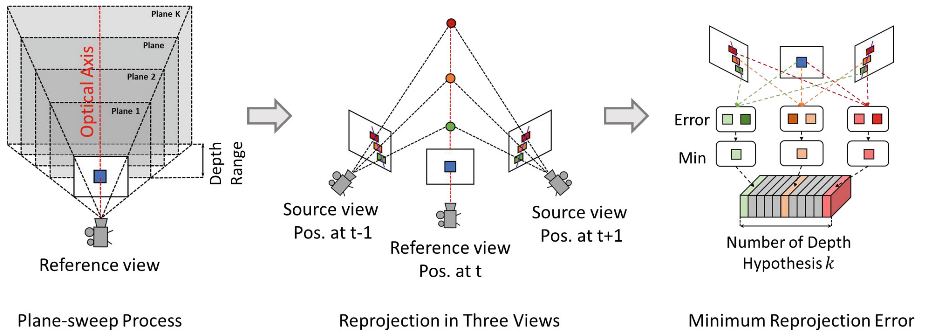

Fig. 2. Visualization depicting cost volume construction. It consists of a plane-sweep process, reprojection in three views as inputs and adopted minimum reprojection error.

image $\mathbf{I}_{\mathrm{s}\to\mathrm{r}}$ through $\mathbf{I}_{\mathrm{s}\to\mathrm{r}}(\boldsymbol{p}_{\mathrm{r}}) = \mathbf{I}_{\mathrm{s}}(\boldsymbol{p}_{\mathrm{s}})$. To implement the self-supervised learning strategy, the reprojection error is calculated based on $\mathbf{I}_{\mathrm{r}}$ and $\mathbf{I}_{\mathrm{s}\to\mathrm{r}}$ by

$$\mathrm{E}(\mathbf{I}_{\mathrm{r}}, \mathbf{I}_{\mathrm{s}\to\mathrm{r}}) = \frac{\alpha}{2}\mathcal{L}_{SSIM}(\mathbf{I}_{\mathrm{r}}, \mathbf{I}_{\mathrm{s}\to\mathrm{r}}) + (1-\alpha)\mathcal{L}_{L1}(\mathbf{I}_{\mathrm{r}}, \mathbf{I}_{\mathrm{s}\to\mathrm{r}}), \tag{2}$$

with

$$\mathcal{L}_{SSIM}(\mathbf{I}_{\mathrm{r}}, \mathbf{I}_{\mathrm{s}\to\mathrm{r}}) = 1 - \mathrm{SSIM}(\mathbf{I}_{\mathrm{r}}, \mathbf{I}_{\mathrm{s}\to\mathrm{r}})) \tag{3}$$

and

$$\mathcal{L}_{L1}(\mathbf{I}_{\mathrm{r}}, \mathbf{I}_{\mathrm{s}\to\mathrm{r}}) = \|\mathbf{I}_{r} - \mathbf{I}_{\mathrm{s}\to\mathrm{r}}\|_1, \tag{4}$$

where structured similarity (SSIM) [24] and L1-norm operator both adopt α at 0.85 followed as [3,13]. Instead of adding the auxiliary task proposed in prior work [12], we back-project the 2D pixel coordinates to the 3D coordinates $\mathrm{P}(\boldsymbol{p}) = \mathbf{K}^{-1}\mathbf{D}(\boldsymbol{p})\boldsymbol{p}$. All the 3D coordinates from $\mathbf{I}_{\mathrm{r}}$ and $\mathbf{I}_{\mathrm{s}}$ gather as point cloud $\mathbf{S}_{\mathrm{r}}$ and $\mathbf{S}_{\mathrm{s}}$. Then synthesized point cloud is warped as $\mathbf{S}_{\mathrm{s}\to\mathrm{r}}(\boldsymbol{p}_{\mathrm{r}}) = \mathbf{S}_{\mathrm{s}}(\boldsymbol{p}_{\mathrm{s}})$. We construct the point cloud consistency by

$$\mathcal{L}_{p} = \mathcal{L}_{L1}(\mathbf{S}_{\mathrm{r}}, \mathbf{S}_{\mathrm{s}\to\mathrm{r}}), \tag{5}$$

where $\mathbf{S}_{\mathrm{r}}$ and $\mathbf{S}_{\mathrm{s}\to\mathrm{r}}$ are based on depth maps from the monocular depth network.

2.2 Improving Depth with Multi-view Guidance

Unlike MDE, MVDE fully leverages the temporal information in the short clip when estimating depth maps. MVDE first samples hypothesized depths value d_k in a depth range from d_{min} to d_{max}. k is the number of sampled depth values. Then, a feature extractor $f_{\mathrm{F}}(\mathbf{I}_{\mathrm{r}}, \mathbf{I}_{\mathrm{s}}; \theta_{\mathrm{F}})$ obtains the deep feature map $\mathbf{F}_{\mathrm{r}}$ and $\mathbf{F}_{\mathrm{s}}$ from input images $\mathbf{I}_{\mathrm{r}}$ and $\mathbf{I}_{\mathrm{s}}$. Similar to the pixel-coordinate matching between $\mathbf{I}_{\mathrm{r}}$ and $\mathbf{I}_{\mathrm{s}}$ as Eq. 1, 2D pixel coordinates of $\mathbf{F}_{\mathrm{s}}$ is back-projected to the each plane

$\mathbf{Z}^{d_k}$ of hypothesised depth value. $\mathbf{Z}^{d_k}$ shares the same depth value d for each pixels. Then the pixel coordinates are matched by

$${}_{\mathrm{f}}\boldsymbol{p}_{\mathrm{s}}^{d} = \mathbf{K}\mathbf{T}_{\mathrm{r}}^{\mathrm{s}}\mathbf{Z}^{d_k}\left({}_{\mathrm{f}}\boldsymbol{p}_{\mathrm{r}}\right)\mathbf{K}^{-1}{}_{\mathrm{f}}\boldsymbol{p}_{\mathrm{r}}, \tag{6}$$

where ${}_{\mathrm{f}}\boldsymbol{p}_{\mathrm{r}}$ is the pixel coordinates in $\mathbf{F}_{\mathrm{r}}$ and ${}_{\mathrm{f}}\boldsymbol{p}_{\mathrm{s}}^{d}$ is the pixel coordinates in $\mathbf{F}_{\mathrm{s}}$ based on the depth value d. Then $\mathbf{F}_{\mathrm{s}}$ is warped to the synthesis feature map by $\mathbf{F}_{\mathrm{s}\to\mathrm{r}}^{d}\left({}_{\mathrm{f}}\boldsymbol{p}_{\mathrm{r}}\right) = \mathbf{F}_{\mathrm{s}}\left({}_{\mathrm{f}}\boldsymbol{p}_{\mathrm{s}}^{d}\right)$. Feature volumes $\mathbf{V}_{\mathrm{r}}$ and $\mathbf{V}_{\mathrm{s}\to\mathrm{r}}$ are aggregations of feature maps $\mathbf{F}_{\mathrm{r}}$ and $\mathbf{F}_{\mathrm{s}\to\mathrm{r}}$. We construct a cost volume $\mathbf{C}$ by

$$\mathbf{C} = \mathcal{L}_{L1}\left(\mathbf{V}_{\mathrm{r}}, \mathbf{V}_{\mathrm{s}\to\mathrm{r}}\right) = \left\|\mathbf{V}_{\mathrm{r}} - \mathbf{V}_{\mathrm{s}\to\mathrm{r}}\right\|_1 . \tag{7}$$

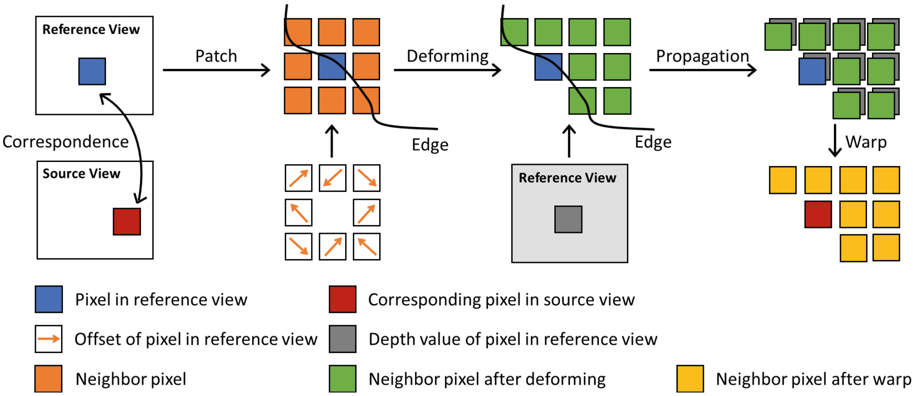
Reference View
Patch
Correspondence
Source View
Deforming
Edge
Propagation
Edge
Reference View
Warp
🟦 Pixel in reference view
🟥 Corresponding pixel in source view
→ Offset of pixel in reference view
⬛ Depth value of pixel in reference view
🟧 Neighbor pixel
🟩 Neighbor pixel after deforming
🟨 Neighbor pixel after warp

Fig. 3. Illustration of deformable patch matching process with pixel coordinates offset and depth propagation.

Previous approaches average the difference between $\mathbf{V}_{\mathrm{r}}$ and all $\mathbf{V}_{\mathrm{s}\to\mathrm{r}}$ from adjacent views to generate cost volumes without considering the occlusion problem and uniform differences between the reference and adjacent feature maps [25,26]. Motivated by these challenges, we introduce the minimum reprojection loss to $\mathbf{C}$, as shown in Fig. 2. We construct the cost volume $\hat{\mathbf{C}}$ via the minor difference value on the corresponding coordinates of the feature volumes as

$$\hat{\mathbf{C}} = \min_{\mathrm{s}} \mathcal{L}_{L1}\left(\mathbf{V}_{\mathrm{r}}, \mathbf{V}_{\mathrm{s}\to\mathrm{r}}\right). \tag{8}$$

We construct a consistency between MVDE and MDE by

$$\mathcal{L}_c = \mathcal{L}_{L1}\left(\mathbf{D}_{\mathrm{r}}^{m}, \mathbf{D}_{\mathrm{r}}^{mv}\right) = \left\|\mathbf{D}_{\mathrm{r}}^{m} - \mathbf{D}_{\mathrm{r}}^{mv}\right\|_1 , \tag{9}$$

where $\mathbf{D}_{\mathrm{r}}^{m}$ and $\mathbf{D}_{\mathrm{r}}^{mv}$ are the depth map predicted by our networks $f_{\mathrm{D}}^{\mathrm{m}}\left(\mathbf{I}_{\mathrm{r}}, \mathbf{I}_{\mathrm{s}}; \theta_{\mathrm{D}}^{\mathrm{m}}\right)$ and $f_{\mathrm{D}}^{\mathrm{mv}}\left(\hat{\mathbf{C}}, \mathbf{F}_{\mathrm{r}}, \mathbf{F}_{\mathrm{s}}; \theta_{\mathrm{D}}^{\mathrm{mv}}\right)$.

2.3 Deformable Patch Matching and Cycled Prediction Learning

Since large areas of textureless areas and reflective parts will cause brightness-based reprojection errors to become unreliable. Furthermore, the homogeneous color on the edge of organs causes local regions to be regarded in the same depth plane. We introduced deformable patch-matching-based local spatial propagation to MDE. As shown in Fig. 3, an offset map $\mathbf{O}_{\mathrm{r}}(\boldsymbol{p}_{\mathrm{r}})$ is adopted to obtain a local region for each pixel by transforming the pixel coordinates in the reference image $\mathbf{I}_{\mathrm{r}}$. Inspired by [15,23], to avoid marginal areas affecting the spatial coherence of the local region, an offset network $f_{\mathrm{O}}\left(\mathbf{I}_{\mathrm{r}};\theta_{\mathrm{O}}\right)$ generates a pixel-level add additional offset map $\Delta\mathbf{O}_{\mathrm{r}}(\boldsymbol{p}_{\mathrm{r}})$. The deformable local region for each pixel can be obtained by

$$\mathbf{R}_{\mathrm{r}} = \boldsymbol{p}_{\mathrm{r}} + \mathbf{O}_{\mathrm{r}}\left(\boldsymbol{p}_{\mathrm{r}}\right) + \Delta\mathbf{O}_{\mathrm{r}}\left(\boldsymbol{p}_{\mathrm{r}}\right), \tag{10}$$

where $\mathbf{R}_{\mathrm{r}}$ is the deformable local regions for pixel $\boldsymbol{p}_{\mathrm{r}}$. After sharing the same depth value by depth propagation in the local region, we implement the deformable local regions on Eq. 1 to complete patch matching by

$$\mathbf{R}_{\mathrm{s}} = \mathbf{K}\mathbf{T}_{\mathrm{r}}^{\mathrm{s}}\mathbf{D}_{\mathrm{r}}\left(\mathbf{R}_{\mathrm{r}}\right)\mathbf{K}^{-1}\mathbf{R}_{\mathrm{r}}, \tag{11}$$

where $\mathbf{R}_{\mathrm{s}}$ is the matched local regions in source views $\mathbf{I}_{\mathrm{s}}$. Based on $\mathbf{R}_{\mathrm{r}}$ and $\mathbf{R}_{\mathrm{s}}$, $\mathbf{I}_{\mathrm{s}}\left(\mathbf{R}_{\mathrm{s}}\right)$ is warped to the synthesised regions $\mathbf{I}_{\mathrm{s}\to\mathrm{r}}\left(\mathbf{R}_{\mathrm{r}}\right)$. The patch-matching-based reprojection error is calculated by

$$\mathcal{L}_{r} = \mathrm{E}\left(\mathbf{I}_{\mathrm{r}}\left(\mathbf{R}_{\mathrm{r}}\right), \mathbf{I}_{\mathrm{s}\to\mathrm{r}}\left(\mathbf{R}_{\mathrm{r}}\right)\right). \tag{12}$$

To better use the temporal correlation, we considered each image as a reference to construct a cycled prediction learning for depth and pose. The total loss on the final computation is averaged from the error of each combination as

$$\mathcal{L}_{total} = \frac{1}{3}\sum_{i=1}^{3}\mathcal{L}_{r}^{m}\left(i\right) + \gamma\mathcal{L}_{r}^{mv}\left(i\right) + \mu\mathcal{L}_{p}\left(i\right) + \lambda\mathcal{L}_{c}\left(i\right) + \delta\mathcal{L}_{s}\left(i\right), \tag{13}$$

where $\mathcal{L}_{r}^{m}$ is the reprojection error term for MDE, and $\mathcal{L}_{r}^{mv}$ is for MVDE. i is the index number of views in the short clip. $\mathcal{L}_{s}$ is the smoothness term [3,12].

3 Experiments

3.1 Datasets and Evaluation Metrics

SCARED [1] datasets were adopted for all the experiments. This dataset contained 35 laparoscopic stereo videos with nine different scenes. And the corresponding depth values obtained through coded structured light images served as ground-truth. We divided the SCARED datasets into a 10:1:1 ratio for each scene based on the video sequence to conduct our experiments. For training, validation, and testing, there were 23,687, 2,405, and 2,405 frames, respectively. Because of limitations in computational resources, we resized the images to 320 × 256 pixels, a quarter of their original dimensions. Following the previous methods [3,25], we adopted seven classical 2D metrics to evaluate the predicted depth maps. Additionally, we only used the monocular depth model to predict depth values with a single RGB image as input during testing.

Table 1. Depth estimation quantitative results. * denotes the method need multi-frame at test. † denotes that the input images are five frames instead of three during training time.

	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Baseline	0.079	0.999	7.868	0.103	0.918	0.995	**1.000**
Monodepth2 [3]	0.083	0.994	8.167	0.107	0.937	0.995	**1.000**
HR-Depth [13]	0.080	0.938	7.943	0.104	0.940	0.996	**1.000**
Manydepth [25]*	0.075	0.830	7.403	0.099	0.945	0.996	**1.000**
MonoVit [29]	0.074	0.865	7.517	0.097	0.949	0.996	**1.000**
SC-Depth [11]†	0.070	0.744	6.932	0.092	0.951	0.998	0.999
AJ-Depth [10]	0.078	0.896	7.578	0.101	0.937	0.996	**1.000**
GCDepthL [12]	0.071	0.801	7.105	0.094	0.938	0.996	**1.000**
Ours	**0.066**	**0.655**	**6.441**	**0.086**	**0.955**	**0.999**	**1.000**

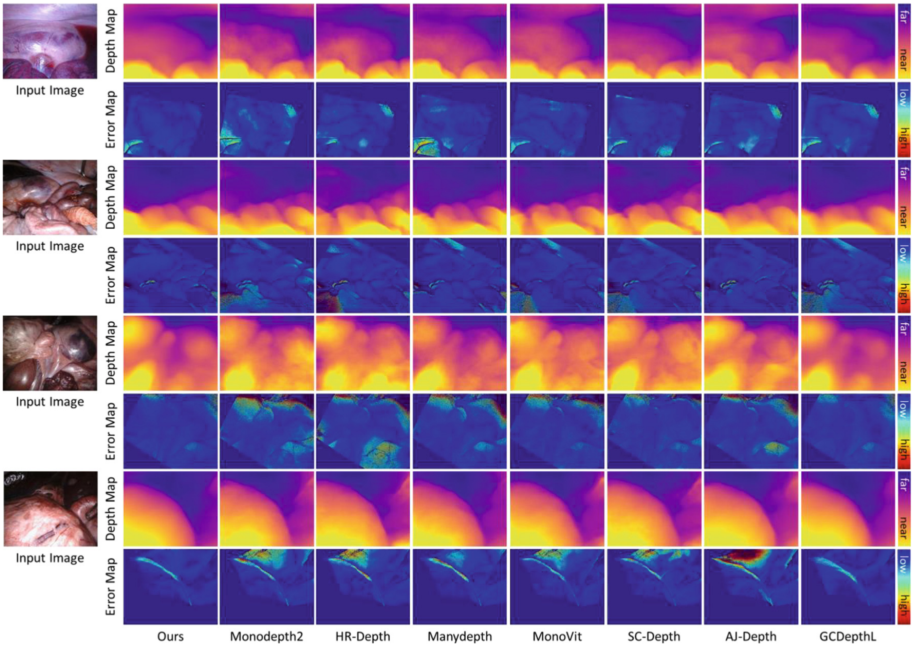

Fig. 4. Depth estimation qualitative results: input images (first column), predicted depth maps and error maps calculated via the absolute relative error metric

3.2 Implementation Details

We utilized PyTorch [16] for model training, employing the Adam optimizer [8] across 25 training epochs. The learning rate started at 1×10^{-4} and dropped by a scale factor of 10 for the final 10 epochs. A batch size is 6, and the total loss function's parameters γ, μ, λ, and δ were set to 0.5, 0.5, 0.2, and 1×10^{-3}, respectively. Additionally, we capped the predicted depth values at 150mm. To construct the cost volume, we adopted the adaptive depth range method [25] and set the number of hypothesized depth values k to 96.

Following [25], we used ResNet-18 [5] with pretrained weights on the ImageNet dataset [19] as encoder module. The feature extractor comprised the first five layers of ResNet-18 [5]. The offset network was two 2D convolution layers.

3.3 Comparison Experiments

We conducted a comprehensive evaluation of our proposed method by comparing it with several classical and state-of-the-art techniques [3,10–13,25,29] retrained on SCARED datasets [1]. Table 1 presents the quantitative results. We also assessed the baseline performance of our proposed method. We compared the depth maps and generated error maps on various laparoscopic scenes based on absolute relative error [25], as shown in Fig. 4.

Table 2. Ablation results for each component contribution in the proposed method. * denotes the method need multi-frame at test. PCC: Point Cloud Consistency; MRE: Minimum Reprojection Error; DPM: Deformable Patch Matching; CPL: Cycled Prediction Learning. M represents monocular depth estimation; MV represents multi-view depth estimation.

	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
w/o PCC	0.084	1.364	9.217	0.116	0.925	0.990	0.997
w/o MRE	0.068	0.711	6.576	0.088	0.953	0.998	0.999
w/o DPM	0.069	0.753	6.901	0.091	0.945	0.997	**1.000**
w/o CPL	0.071	0.702	6.868	0.091	0.956	0.998	**1.000**
Baseline (M) [12]	0.071	0.801	7.105	0.094	0.938	0.996	**1.000**
w/ PCC	0.070	0.763	6.959	0.092	0.948	0.997	**1.000**
Baseline (MV) [25]*	0.075	0.830	7.403	0.099	0.945	0.997	**1.000**
w/ MRE*	0.074	0.797	7.241	0.095	0.953	0.997	**1.000**
Ours (M)	0.068	0.708	6.788	0.089	**0.955**	0.998	**1.000**
Ours (MV)*	0.070	0.743	6.842	0.091	0.948	0.997	**1.000**
Ours	**0.066**	**0.655**	**6.441**	**0.086**	**0.955**	**0.999**	**1.000**

3.4 Ablation Study

We conducted an ablation study to evaluate the influence of different components in our proposed approach. Table 2 shows the results of our method with four different components, namely, point cloud consistency (PCC), minimum reprojection error (MRE), deformable patch matching (DPM), and cycled prediction learning (CPL). We adopted GCDepthL [12] and Manydepth [25] as baselines for our method's monocular and multi-view depth models. We proposed PCC and MRE as two optimized modules for these baseline models and evaluated their impact on each baseline individually. Furthermore, we trained the monocular and multi-view depth models separately in our proposed method without consistency to demonstrate the contribution of combining these two models.

4 Discussion and Conclusions

The laparoscopic scenes typically feature large, smooth regions with organ surfaces and homogeneous colors along the edges of organs. This can cause issues while matching pixels, as the points in the boundary area can be mistaken to be in the same depth plane. Our proposed method's depth maps, as shown in Fig 4, exhibit a smoother performance in the large regions of input images when compared to the existing methods [3,10–13,25,29]. Additionally, the error maps reveal that our proposed method performs better even when the depth maps look similar qualitatively. At the marginal area of organs, our proposed method generates better depth predictions with smoother depth maps and lower errors, despite the color changes being barely perceptible with the depth value changes. Our proposed method outperforms current approaches on the seven metrics we used, as demonstrated in Table 1. The ablation study reveals that the proposed method improves significantly when combining each component, and each component contributes to the proposed method. Specifically, the optimized modules PCC and MRE designed for monocular and multi-view depth models enhance the performance of the baselines [12,25]. The combination of monocular and multi-view depth models yields better results than the single model trained independently, as seen in the last three rows of Table 2.

In conclusion, we incorporate more temporal information in the monocular depth model by leveraging the guidance of the multi-view depth model when predicting depth values. We introduce the minimum reprojection error to construct the multi-view depth model's cost volume and optimize the monocular depth model's point cloud consistency module. Moreover, we propose a novel method that matches deformable patches in spatially coherent local regions instead of point matching. Finally, cycled prediction learning is designed to exploit temporal information. The outcomes of the experiments indicate an improved depth estimation performance using our approach.

Acknowledgments. We extend our appreciation for the assistance provided through the JSPS Bilateral International Collaboration Grants; JST CREST Grant (JPMJCR20D5); MEXT/JSPS KAKENHI Grants (17H00867, 26108006, 21K19898); and the CIBoG initiative of Nagoya University, part of the MEXT WISE program.

References

1. Allan, M., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems 27 (2014)

3. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3828–3838 (2019)
4. Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. arXiv preprint arXiv:2002.12319 (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitionm, pp. 770–778 (2016)
6. Huang, B., et al.: Self-supervised generative adversarial network for depth estimation in laparoscopic images. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 227–237. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_22
7. Hwang, M., et al.: Applying depth-sensing to automated surgical manipulation with a da Vinci robot. In: 2020 International Symposium on Medical Robotics (ISMR), pp. 22–29. IEEE (2020)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 239–248. IEEE (2016)
10. Li, W., Hayashi, Y., Oda, M., Kitasaka, T., Misawa, K., Kensaku, M.: Attention guided self-supervised monocular depth estimation based on joint depth-pose loss for laparoscopic images. Comput. Assist. Radiol. Surg. (2022)
11. Li, W., Hayashi, Y., Oda, M., Kitasaka, T., Misawa, K., Mori, K.: Spatially variant biases considered self-supervised depth estimation based on laparoscopic videos. Comput. Methods Biomech. Biomed. Eng.: Imaging Vis., 1–9 (2021)
12. Li, W., Hayashi, Y., Oda, M., Kitasaka, T., Misawa, K., Mori, K.: Geometric constraints for self-supervised monocular depth estimation on laparoscopic images with dual-task consistency. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, LNCS, Part IV, pp. 467–477. Springer (2022). https://doi.org/10.1007/978-3-031-16440-8_45
13. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: HR-Depth: high resolution self-supervised monocular depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2294–2301 (2021)
14. Ming, Y., Meng, X., Fan, C., Yu, H.: Deep learning for monocular depth estimation: a review. Neurocomputing **438**, 14–33 (2021)
15. Park, J., Joo, K., Hu, Z., Liu, C.-K., So Kweon, I.: Non-local spatial propagation network for depth completion. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12358, pp. 120–136. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58601-0_8
16. Paszke, A., et al.: Automatic differentiation in PyTorch. In: NIPS 2017 Workshop on Autodiff (2017)
17. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3227–3237 (2020)
18. Qian, L., Zhang, X., Deguet, A., Kazanzides, P.: ARAMIS: augmented reality assistance for minimally invasive surgery using a head-mounted display. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11768, pp. 74–82. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32254-0_9

19. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
20. Sánchez-González, P., et al.: Laparoscopic video analysis for training and image-guided surgery. Minim. Invasive Therapy Allied Technol. **20**(6), 311–320 (2011)
21. Tosi, F., Aleotti, F., Poggi, M., Mattoccia, S.: Learning monocular depth estimation infusing traditional stereo knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9799–9809 (2019)
22. Vecchio, R., MacFayden, B., Palazzo, F.: History of laparoscopic surgery. Panminerva Med. **42**(1), 87–90 (2000)
23. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: learned multi-view patchmatch stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14194–14203 (2021)
24. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
25. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: 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 (2021)
26. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: MVSNet: depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 767–783 (2018)
27. Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.Z.: Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint arXiv:1705.08260 (2017)
28. Zhao, C., Yen, G.G., Sun, Q., Zhang, C., Tang, Y.: Masked GAN for unsupervised depth and pose prediction with scale consistency. IEEE Trans. Neural Netw. Learn. Syst. **32**(12), 5392–5403 (2020)
29. Zhao, C., et al.: MonoViT: self-supervised monocular depth estimation with a vision transformer. arXiv preprint arXiv:2208.03543 (2022)
30. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851–1858 (2017)