Enforcing Geometric Constraints of Surface Normal and Pose for Self-supervised Monocular Depth Estimation on Laparoscopic Images

Wenda Li[1][0000-0003-4486-7207], Yuichiro Hayashi[1],
Masahiro Oda[2,1][0000-0001-7714-422X], Takayuki Kitasaka[3],
Kazunari Misawa[4], and Kensaku Mori[1,2,5][0000-0002-0100-4797]

[1] Graduate School of Informatics, Nagoya University,
Furou-cho, Chikusa-ku, Nagoya, 464-8601, Aichi, Japan
[2] Information Technology Center, Nagoya University,
Furou-cho, Chikusa-ku, Nagoya, 464-8601, Aichi, Japan
[3] Faculty of Information Science, Aichi Institute of Technology,
Yakusacho, Toyota, 470-0392, Aichi, Japan
[4] Aichi Cancer Center Hospital, Chikusa-ku, Nagoya, 464-8681, Aichi, Japan
[5] Research Center of Medical Bigdata, National Institute of Informatics,
Hitotsubashi, Chiyoda-ku, 101-8430, Tokyo, Japan

Abstract. Depth information is essential for 3D reconstruction in surgical scenes. Depth-pose-based self-supervised monocular depth estimation has advanced significantly but faces two challenges in laparoscopic scenes, leading to unreliable pixel matching during training. This also results in depth maps failing to preserve geometric structure when back-projected into 3D space. Second, limited movement space necessitates that laparoscopic motion involves pure complex rotations. It further complicates the relative pose estimation between adjacent views. To address these issues, we propose a novel self-supervised monocular depth estimation method guided by geometric constraints. We incorporate surface normal estimation with depth-normal consistency to establish a geometric constraint for predicted depth maps. Furthermore, we propose an uncertainty measure based on the distance from 3D points to a synthesized plane, reducing conversion bias from depth to normals. Moreover, we optimize pose estimation using a feature-matching process with a 4D score volume. Our method reduced absolute relative error by 19.0% and 3D completeness by 23.9% over the baseline. Our code is available at https://github.com/MoriLabNU/GSPDepthL.

Keywords: Depth estimation · Self-supervised learning · Laparoscopy.

1 Introduction

Depth information plays a crucial role in mapping surgical fields in robotic-assisted minimally invasive surgery (RAMIS) and augmented-reality-assisted minimally invasive surgery (ARAMIS) [7, 14, 16]. It is used to create realistic

3D scenes and 3D models for surgical navigation and surgeon training. Monocular depth estimation (MDE) predicts a pixel-level depth map from a single image, and learning-based approaches have seen significant advancements.

Supervised learning requires extensive annotated data with complex network architectures, which is both costly in terms of time and resources for depth value collection and training [17]. To address this, researchers explored self-supervised methods for MDE. Zhou et al. [26] first introduced depth-pose self-supervised MDE, completing a pixel matching process between adjacent images based on predicted relative poses. Godard et al. [3] optimized this work by minimizing reprojection error and established a widely used baseline. Subsequent innovations in self-supervised MDE have included approaches that leverage segmentation [4], multi-frame constraint [19] and transformer [24]. Recently, self-supervised MDE has been developed for laparoscopic scenes. Huang et al. [6] leveraged 3D points for self-supervised MDE using stereo datasets rather than monocular scenes. Shao et al. [18] proposed an appearance module to realize brightness consistency for laparoscopic scenes. Li et al. [11] employed block matching instead of pixel matching to improve depth estimation on smooth surfaces. Cui et al. [2] first introduced foundation models for self-supervised MDE.

Existing self-supervised MDE faces two key challenges in laparoscopic scenes. First, homogeneous textures and colors on organ surfaces reduce photometric error, even with incorrect pixel matching. It also leads to depth maps that lack geometric structure when back-projected depths to 3D space. Second, due to the limited space for movement, the laparoscope's motions involve many pure complex rotations. Pure complex rotations without translations increase the difficulty of pose estimation. As an earlier work, Yang et al. [22] proposed edge-aware depth-normal consistency for autonomous and indoor scenes, but the smoothness of laparoscopic images makes edge detection more difficult. AF-SfMLearner [18] and MGMNet [11] improved depth predictions on smooth surfaces but overlook the geometric structure of the estimated depth maps. GCDepthL [10] enforced consistency between the predicted scene coordinates and depth maps, applying a per-point constraint highly susceptible to noise. In addition, many previous methods [13, 18, 11] did not explicitly address pose estimation, despite its challenges in laparoscopic datasets. GCDepthL [10] optimized pose estimation similarly to stereo matching, but pure complex rotations in laparoscopic scenes further complicate the process. To address these challenges, we introduce a depth-normal consistency framework with a novel distance-based uncertainty mechanism, thereby enhancing the robustness of depth estimation. This consistency enforces smoothness of local depth variations through a depth-to-normal transformation while maintaining global geometric constraints. In addition, we incorporate feature-matching into the pose estimation through a 4D score volume. This approach leveraged the spatial information between the feature maps extracted from adjacent images.

Our main contributions are summarized as follows. (i) We introduce surface normal estimation and build the depth-normal consistency to guide monocular depth estimation and provide geometric constraints. (ii) We model an uncertainty map for the depth-normal consistency to alleviate bias when converting the depths to normal vector in a local region. (iii) We propose a feature-matching process by calculating the 4D score volume to optimize the pose estimation.

2 Method

2.1 Self-supervised MDE with 4D Score Volume

Following the previous method [3], we consider the self-supervised MDE as the view-synthesis problem. As shown in Fig. 1, the inputs of the whole network are target image $\mathbf{I}_t$ from the view at time t and source images $\mathbf{I}_s$ from the adjacent view at time s. Time s is time $t-1$ or time $t+1$.

As shown in Fig. 2 (a), previous methods [10, 18] estimate the relative pose without considering spatial information between adjacent images. Therefore, we introduce a 4D score volume into the pose estimation network. As shown in Fig. 2 (b), we firstly use the feature extractor to obtain the feature maps $\mathbf{F}_t$ and $\mathbf{F}_s$ from the target image $\mathbf{I}_t$ and source images $\mathbf{I}_s$. Then, we calculate the 4D score volume to implicitly complete feature-matching based on $\mathbf{F}_t$ and $\mathbf{F}_s$ by

$$\mathbf{V}^{\boldsymbol{p}_c,\boldsymbol{q}_c} = \frac{\sum_C \mathbf{F}_t^{\boldsymbol{p}_c} \cdot \mathbf{F}_s^{\boldsymbol{q}_c}}{\sqrt{\sum_C \left(\mathbf{F}_t^{\boldsymbol{p}_c}\right)^2} \cdot \sqrt{\sum_C \left(\mathbf{F}_s^{\boldsymbol{q}_c}\right)^2}}, \tag{1}$$

where $\boldsymbol{p}_c$ and $\boldsymbol{q}_c$ are pixels' 2D coordinates on the channel with index c of feature maps $\mathbf{F}_t$ and $\mathbf{F}_s$. C is the number of channels of a feature map. Then we input $\mathbf{F}_t$ and $\mathbf{F}_s$ and normalized 4D score volume $\mathbf{V}$ to pose estimation network. The output is the transformation matrix $\mathbf{T}_{t\to s}$, which represents the relative pose of the laparoscope from the view at time t to the adjacent view at time s.

The predicted transformation matrix $\mathbf{T}_{t\to s}$ and the laparoscope's intrinsic parameters $\mathbf{K}$ are used for pixel matching between the target image $\mathbf{I}_t$ and source images $\mathbf{I}_s$. Given a pixel at $\mathbf{p}_t$ in $\mathbf{I}_t$, its corresponding coordinate in $\mathbf{I}_s$ is computed as:

$$\boldsymbol{p}_s = \mathbf{K}\mathbf{T}_{t\to s}\mathbf{D}_t^{\boldsymbol{p}_t}\mathbf{K}^{-1}\boldsymbol{p}_t, \tag{2}$$

where $\mathbf{D}_t^{\boldsymbol{p}t}$ is the depth at $\mathbf{p}_t$ in $\mathbf{D}_t$, and $\mathbf{K}$ is a 4×4 intrinsic matrix. The synthesized frame $\mathbf{I}_{s\to t}$ is obtained by warping $\mathbf{I}_s$, and the difference between $\mathbf{I}_t$ and $\mathbf{I}_{s\to t}$ serves as a supervision signal. Following Monodepth2 [3], the photometric error is defined as:

$$\mathrm{E}\left(\mathbf{I}_t, \mathbf{I}_{s\to t}, \boldsymbol{p}\right) = \alpha\frac{1-\mathrm{SSIM}\left(\mathbf{I}_t, \mathbf{I}_{s\to t}, \boldsymbol{p}\right)}{2} + (1-\alpha)\left\|\mathbf{I}_t^{\boldsymbol{p}} - \mathbf{I}_{s\to t}^{\boldsymbol{p}}\right\|_1, \tag{3}$$

where $\mathbf{I}_t^{\boldsymbol{p}}$ and $\mathbf{I}_{s\to t}^{\boldsymbol{p}}$ is the value at the 2D coordinate $\boldsymbol{p}$ in $\mathbf{I}_t$ and $\mathbf{I}_{s\to t}$. SSIM is local structural similarity, and α is set to 0.85. The minimum reprojection error is defined as:

$$\mathcal{L}_r = \frac{1}{|\mathbf{H}|}\sum_{\boldsymbol{p}\in\mathbf{H}} \min_s \mathrm{E}\left(\mathbf{I}_t, \mathbf{I}_{s\to t}, \boldsymbol{p}\right), \tag{4}$$

where $\boldsymbol{p}$ is the 2D coordinate of a pixel. $\mathbf{H}$ is a set including all pixels' 2D coordinates. $|\cdot|$ represents Cardinal function.

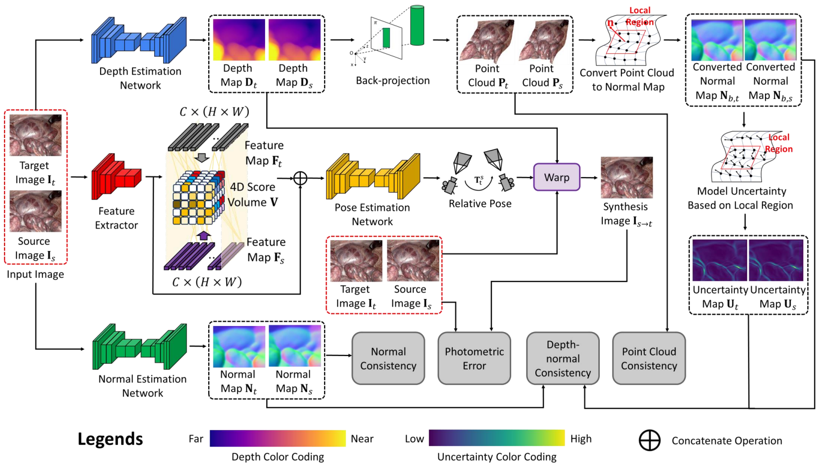
Depth Estimation Network
Depth Map $\mathbf{D}_t$
Depth Map $\mathbf{D}_s$
Back-projection
Point Cloud $\mathbf{P}_t$
Point Cloud $\mathbf{P}_s$
Local Region
n
Convert Point Cloud to Normal Map
Converted Normal Map $\mathbf{N}_{b,t}$
Converted Normal Map $\mathbf{N}_{b,s}$
$C \times (H \times W)$
Feature Map $\mathbf{F}_t$
4D Score Volume $\mathbf{V}$
Feature Map $\mathbf{F}_s$
$C \times (H \times W)$
Target Image $\mathbf{I}_t$
Source Image $\mathbf{I}_s$
Input Image
Feature Extractor
Pose Estimation Network
$\mathbf{T}_t^s$
Relative Pose
Warp
Synthesis Image $\mathbf{I}_{s \to t}$
Local Region
Model Uncertainty Based on Local Region
Uncertainty Map $\mathbf{U}_t$
Uncertainty Map $\mathbf{U}_s$
Target Image $\mathbf{I}_t$
Source Image $\mathbf{I}_s$
Normal Estimation Network
Normal Map $\mathbf{N}_t$
Normal Map $\mathbf{N}_s$
Normal Consistency
Photometric Error
Depth-normal Consistency
Point Cloud Consistency
Legends
Far
Near
Depth Color Coding
Low
High
Uncertainty Color Coding
Concatenate Operation

Fig. 1. Overview of our self-supervised monocular depth estimation framework. The proposed method consists of a monocular depth estimation network, a pose estimation network, and a normal estimation network. Normal map includes the components of the normal vector in three channels. C, H and W are the channel number, height and width of feature maps. 4D score volume has a size of $H \times W \times H \times W$.

2.2 Depth-normal Consistency under Distance-based Uncertainty

To enforce geometric constraints in depth estimation, we introduce surface normal estimation and establish depth-normal consistency, as shown in Fig. 1. A normal map encodes surface normals at each 3D point, computed by fitting a local plane to its neighbors. This consistency enforces smooth depth variations by converting depth values into surface normals based on local regions. We then back-project the estimated depth map $\mathbf{D}$ into 3D space using the laparoscope's intrinsic matrix $\mathbf{K}$ by

$$\mathbf{P}^{\boldsymbol{p}_t} = \mathbf{K}^{-1}\mathbf{D}_t^{\boldsymbol{p}_t}\boldsymbol{p}_t, \tag{5}$$

where $\mathbf{P}$ is the point cloud consisting of 3D points. $\mathbf{P}^{\boldsymbol{p}_t}$ is back-projected 3D points corresponding to the 2D coordinates $\boldsymbol{p}_t$ in $\mathbf{D}_t$. As shown in Fig. 3 (a), we convert the back-projected 3D points to normal map $\mathbf{N}_b$ by

$$\mathbf{N}_b^{\boldsymbol{p}} = \frac{1}{|\boldsymbol{\Omega}|} \sum_{\boldsymbol{p}^i, \boldsymbol{p}^j \in \boldsymbol{\Omega}} \frac{\left(\mathbf{P}^{\boldsymbol{p}^i} - \mathbf{P}^{\boldsymbol{p}}\right) \times \left(\mathbf{P}^{\boldsymbol{p}^j} - \mathbf{P}^{\boldsymbol{p}}\right)}{\left\|\left(\mathbf{P}^{\boldsymbol{p}^i} - \mathbf{P}^{\boldsymbol{p}}\right) \times \left(\mathbf{P}^{\boldsymbol{p}^j} - \mathbf{P}^{\boldsymbol{p}}\right)\right\|_2}, \tag{6}$$

where $\boldsymbol{\Omega}$ is the set of 2D coordinates of eight surrounding pixels centering on the pixel with 2D coordinate $\boldsymbol{p}$. $\boldsymbol{p}^i$ and $\boldsymbol{p}^j$ are different surrounding pixels' 2D coordinates, belonging to set $\boldsymbol{\Omega}$. × is the cross product. $\mathbf{N}$ is normal maps consist of normal vectors $\boldsymbol{n}$.

The assumption of conversion from 3D points into a normal vector is that all points within a local region lie on the same plane. However, this assumption is

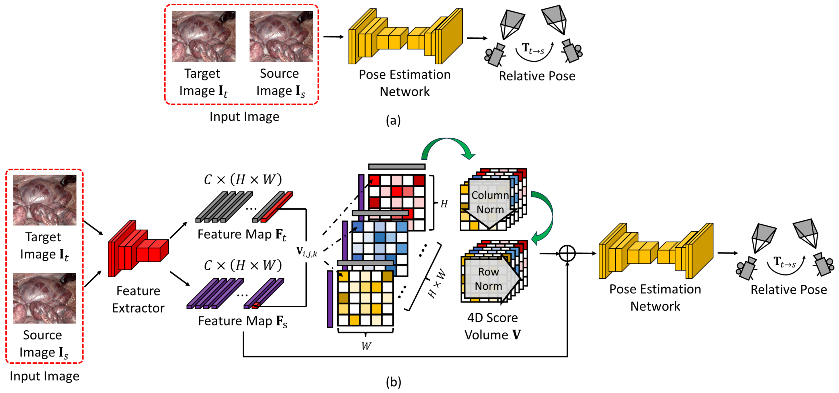

Fig. 2. Pose Estimation Process. (a) Previous pose estimation process. (b) Proposed pose estimation process with 4D score volume based on feature-matching.

compromised due to object boundaries and inaccurate predictions. As shown in Fig. 3 (b), we model an uncertainty map based on the distances of points within the local region to the synthesized plane to mitigate this impact by

$$\mathbf{U}^{\boldsymbol{p}} = \frac{1}{|\boldsymbol{\Omega}|}\sum_{\boldsymbol{p}^k\in\boldsymbol{\Omega}} \left\| \mathbf{N}_b^{\boldsymbol{p}} \cdot \left(\mathbf{P}^{\boldsymbol{p}^k} - \mathbf{P}^{\boldsymbol{p}}\right)\right\|_2, \tag{7}$$

where $\mathbf{U}$ is the uncertainty map. $\boldsymbol{p}^k$ are 2D coordinates of surrounding pixels, belonging to set $\boldsymbol{\Omega}$. As shown in Fig. 1, we adopt a normal estimation network to predict normal maps $\mathbf{N}_t$ and $\mathbf{N}_s$ from target image $\mathbf{I}_t$ and source image $\mathbf{I}_s$. The loss function for the depth-normal consistency is defined by

$$\mathcal{L}_c = \frac{1}{|\mathbf{H}|}\sum_{\boldsymbol{p}\in\mathbf{H}} \left(1-\mathbf{U}_t^{\boldsymbol{p}}\right)\left(1-\mathbf{N}_{b,t}^{\boldsymbol{p}}\cdot\mathbf{N}_t^{\boldsymbol{p}}\right) + \left(1-\mathbf{U}_s^{\boldsymbol{p}}\right)\left(1-\mathbf{N}_{b,s}^{\boldsymbol{p}}\cdot\mathbf{N}_s^{\boldsymbol{p}}\right), \tag{8}$$

where $\boldsymbol{p}$ is 2D coordinates of pixel in the normal map. $\mathbf{H}$ is a set that includes all pixels' 2D coordinates in the normal map. $\mathbf{N}_{b,t}$ and $\mathbf{N}_{b,s}$ are the normal maps converted from the predicted depth maps $\mathbf{D}_t$ and $\mathbf{D}_s$. $\mathbf{U}_t$ and $\mathbf{U}_s$ are uncertainty maps based on $\mathbf{N}_t^d$ and $\mathbf{N}_s^d$. $\cdot$ is the dot product. We adopt a negative cosine loss [17] as Eq. 8 for normal supervision and minimize the distance of 3D points from adjacent views as normal by

$$\mathcal{L}_n = \frac{1}{|\mathbf{H}|}\sum_{\boldsymbol{p}\in\mathbf{H}} \left(1-\mathbf{N}_{s\to t}^{\boldsymbol{p}}\cdot\mathbf{N}_t^{\boldsymbol{p}}\right), \tag{9}$$

where $\mathbf{N}_{s\to t}$ is the synthesized normal map and transformed 3D points from $\mathbf{N}_s$ based on the pixel-matching process and coordinate system transformation. Our final loss is $\mathcal{L}_f = \mathcal{L}_r + \lambda\mathcal{L}_c + \gamma\mathcal{L}_n + \mu\mathcal{L}_s + \xi\mathcal{L}_p$, where $\mathcal{L}_s$ is the smoothness term for the predicted depth maps [3] and $\mathcal{L}_p$ is the point cloud consistency as MGMNet [11].

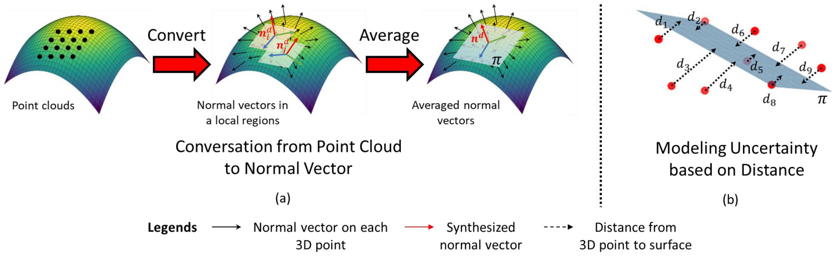
Convert
Average
Point clouds
Normal vectors in
a local regions
Averaged normal
vectors
Conversation from Point Cloud
to Normal Vector
(a)
n_i^d
n_j^d
n^d
π
d_1 d_2 d_3 d_4 d_5 d_6 d_7 d_8 d_9
Modeling Uncertainty
based on Distance
(b)
Legends
Normal vector on each 3D point
Synthesized normal vector
Distance from 3D point to surface

Fig. 3. Visualization depicting (a) the conversion from 3D points to normal vector and (b) the modeling of the uncertainty based on distance. (a) We use the various combinations of surrounding points to obtain normal vectors as $\boldsymbol{n}_i^d$ and $\boldsymbol{n}_j^d$. And average them as the final normal vector $\boldsymbol{n}^d$ corresponding to the center 3D points. (b) We model the uncertainty based on the distance of 3D points to synthesized plane π from (a).

3 Experiments and Results

3.1 Datasets and Evaluation Metrics

We conducted all experiments on the SCARED [1] and Hamlyn datasets [15]. SCARED consists of nine laparoscopic scenes, and we followed the dataset splits used in AF-SfMLearner [18]. Hamlyn datasets provides laparoscopic videos processed by Recasens et al. [15] to create ground truth. Following Monodepth2 [3], we split the Hamlyn dataset into a training set (21,090 frames) and a testing set (2,014 frames), ensuring that the training and testing scenes are distinct. All images were resized to 320×256 due to computational constraints. We evaluated depth predictions using three 2D metrics [3] and assessed back-projected 3D points with two 3D metrics [11]. During testing, only the depth estimation network was used, taking a single image as input.

3.2 Implementation Details

We re-trained all models by PyTorch with the Adam optimizer [8] for 30 epochs. The learning rate was set at 1×10^{-4}, with a reduction by a factor of 10 after 15 epochs. The training utilized a batch size of 12, and the total loss function parameters, λ, γ, μ, and ξ were designated as 0.01, 0.01, 0.001, and 0.001. Furthermore, we cap the depth values at 150 mm and 180 mm for SCARED [1] and Hamlyn [15]. The model was conducted on an NVIDIA Quadro RTX 6000 GPU for 15 hours. Following Monodepth2 [3], all encoder modules incorporated a ResNet-18 with pre-trained weights from the ImageNet dataset [3]. Our decoder followed the design outlined in Monodepth2 [3].

3.3 Comparison Evaluation

We compared the proposed method with several existing approaches [2, 3, 5, 9–13, 18–21, 23–25]. We retrained them three times with different seeds on SCARED [1]

Table 1. Quantitative comparison for predicted depths and back-projected 3D points on SCARED and Hamlyn. The best results are bold. The second-best results are underlined. * denotes the backbone is foundation models as EndoDAC [2].

Datasets	SCARED					Hamlyn				
Metrics	2D Metrics			3D Metrics		2D Metrics			3D Metrics	
	Abs Rel↓	RMSE↓	$\gamma < 1.25$↑	Comp.↓	Recall↑	Abs Rel↓	RMSE↓	$\gamma < 1.25$↑	Comp.↓	Recall↑
Monodepth2 [3]	0.076	6.127	0.942	3.755	0.646	0.176	16.406	0.754	7.137	0.340
HRDepth [12]	0.072	5.787	0.950	3.364	0.672	0.171	15.627	0.755	6.696	0.347
Manydepth [19]	0.070	5.787	0.957	3.363	0.669	0.166	15.477	0.764	6.607	0.362
DIFFNet [25]	0.066	5.702	0.958	3.085	0.687	0.168	15.409	0.759	6.629	0.344
MonoViT [24]	0.070	5.707	0.954	3.437	0.688	0.175	16.254	0.752	6.955	0.350
BRNet [5]	0.074	6.367	0.943	3.501	0.649	0.163	15.013	0.768	6.476	0.348
Lite-Mono [23]	0.070	6.133	0.951	3.246	0.680	0.169	15.336	$\underline{0.772}$	6.654	0.351
SCDepth [9]	0.071	5.788	0.950	3.461	0.674	0.184	16.961	0.724	6.981	0.330
Endo-SLAM [13]	0.064	5.743	$\underline{0.959}$	2.776	0.733	0.181	16.229	0.704	6.742	0.307
AF-SfMLearner [18]	0.066	$\underline{5.608}$	0.957	3.234	0.703	0.169	15.862	0.739	6.577	0.342
GCDepthL [10]	$\underline{0.062}$	5.851	0.958	$\underline{2.625}$	$\underline{0.729}$	0.162	14.762	0.749	$\underline{5.881}$	$\underline{0.399}$
MGMNet [11]	0.063	5.696	0.958	2.798	0.717	$\underline{0.159}$	$\underline{14.553}$	0.770	6.114	0.369
Baseline	0.068	6.562	0.951	3.117	0.703	0.172	16.703	0.639	6.940	0.343
Ours	**0.055**	**4.800**	**0.969**	**2.435**	**0.777**	**0.143**	**13.142**	**0.795**	**5.202**	**0.437**
EndoDAC [2]	$\underline{0.054}$	$\underline{4.546}$	$\underline{0.976}$	$\underline{2.442}$	$\underline{0.767}$	0.159	13.047	$\underline{0.776}$	7.129	0.337
DA [20]*	0.059	4.980	0.967	2.748	0.740	0.161	12.693	0.744	6.418	$\underline{0.348}$
DAV2 [21]*	0.079	6.413	0.942	3.777	0.615	$\underline{0.158}$	$\underline{12.643}$	0.750	$\underline{6.391}$	0.340
Ours*	**0.049**	**4.022**	**0.981**	**2.045**	**0.805**	**0.127**	**10.076**	**0.810**	**5.005**	**0.451**

and Hamlyn [15] and reported the mean of the results in Tables 1 and 2. Table 1 presents the quantitative results for depth maps and back-projected 3D points on 2D and 3D metrics. The baseline of the proposed method is Monodepth2 [3] with the Siamese pose process proposed by GCDepthL [10]. In addition, we evaluated depth prediction accuracy in different laparoscopic scenes using error maps based on absolute relative error [19], as shown in Fig. 4. Furthermore, we tested our method with the same foundation model as EndoDAC[2] and compared it against other foundation-based methods [2, 20, 21].

3.4 Ablation Study

We performed the ablation study based on ten 2D and 3D metrics for depth estimation and back-projected 3D points to analyze the impact of the components in the proposed method. Table 2 presents the results of our method, which incorporates four proposed components: surface normal, distance uncertainty, normal loss, and 4D cost volume. Since the surface normal was introduced to build the depth-normal consistency, the component named surface normal also included depth-normal loss. Normal loss is denoted as normal consistency loss.

4 Discussion and Conclusion

Due to the characteristics of laparoscopic scenes, previous self-supervised MDE methods perform poorly on laparoscopic datasets. As shown in Table 1, existing methods designed for autonomous driving and indoor datasets [3, 5, 12, 19, 23–25, 20, 21] struggle in laparoscopic scenes compared to laparoscopic-specific

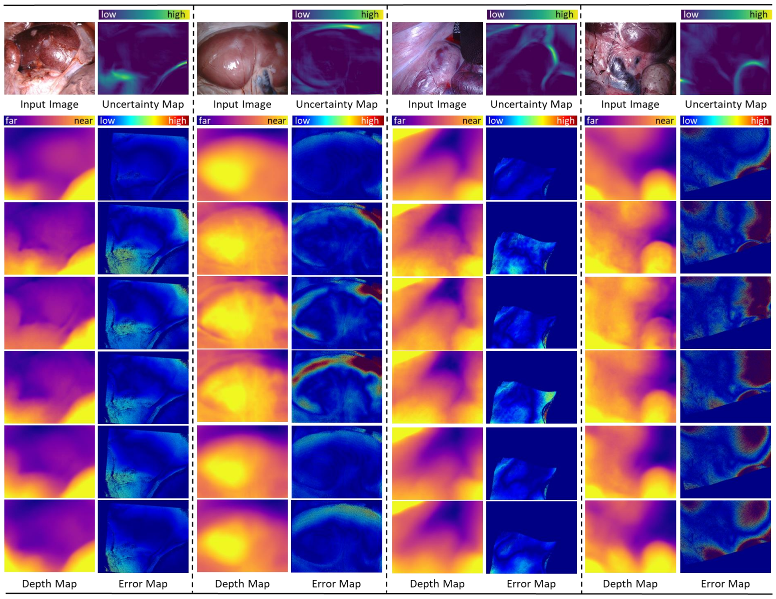

Fig. 4. Comparison of qualitative results for depth estimation. Row 1 shows the input images and the obtained uncertainty map. Rows 2 through 6 depict the estimated depth maps and the error maps calculated by the absolute relative error metric.

foundation models [2, 9–11, 13, 18]. However, prior laparoscopic approaches [2, 9–11, 13, 18] also exhibit limitations, particularly in 3D metrics based on the 3D points back-projected from the predicted depth maps. As show in Table 1, the proposed method not only enhances depth prediction accuracy outperforms existing methods in 3D evaluation on the SCARED and Hamlyn datasets, as shown in Table 1. Notably, despite not using a foundation model as a backbone, the proposed method surpasses foundation-based approaches [2, 20, 21] in 3D metrics. When adopting the same foundation model as EndoDAC [2], the proposed method further improves both 2D and 3D performance compared to other foundation-based approaches [2, 20, 21]. As shown in Table 2, each proposed component contributes to performance improvements (IDs 1-5). And the full model achieved the best results (IDs 1-6). The ablation study shows that the 4D cost volume enhances performance compared to the baseline (IDs 1 and 4). And it also highlights that surface normal estimation with distance uncertainty and normal loss, plays a more significant role in 3D metrics compared to 4D cost volume (IDs 4 and 5). And in the surface normal consistency component, the distance uncertainty is the primary contributing factor (IDs 2-4). The proposed method produces smoother depth maps with lower errors compared to existing methods [3, 10, 11, 18, 24], as shown in Fig. 4. The uncertainty maps indicate higher uncertainty at object edges because the depth-to-normal conversion assumes that the 3D points are in the same plane in a local region.

Table 2. Ablation study with 2D and 3D metrics on SCARED. The best results are bold. The second-best results are underlined.

ID	Components				2D Metrics			3D Metrics	
	Surface Normal	Distance Uncertainty	Normal Loss	4D Cost Volume	Abs Rel ↓	RMSE ↓	$\gamma < 1.25$ ↑	Comp. ↓	Recall ↑
1					0.067	5.539	0.952	3.117	0.703
2	✓				0.061	5.228	0.957	2.796	0.745
3	✓	✓			0.058	4.986	0.964	2.625	0.764
4	✓	✓	✓		0.057	4.881	0.967	2.558	0.768
5				✓	0.061	5.136	0.958	2.840	0.738
6	✓	✓	✓	✓	**0.055**	**4.800**	**0.969**	**2.435**	**0.777**

In conclusion, we analyze the challenges of applying self-supervised MDE to laparoscopic images. We introduce the surface normal estimation and propose a consistency between the predicted depths and the surface normal. We also optimize pose estimation with a 4D score volume based on the feature maps extracted from adjacent images. Experimental results demonstrate that the proposed method had superior 2D and 3D performance, with smoother depth maps and lower errors compared to existing methods. Further efforts will focus on addressing not only smooth regions but also highly folded organs, overlapping structures, and other complex anatomical surfaces.

Acknowledgments. The authors are grateful for support from the JST CREST Grant Number JPMJCR20D5; the MEXT/JSPS KAKENHI Grant Numbers 24H00720, 17H00867, 26108006, and 21K19898; the JSPS Bilateral International Collaboration Grants; and the CIBoG program of Nagoya University with funds from the MEXT WISE program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. Cui, B., Islam, M., Bai, L., Wang, A., Ren, H.: EndoDAC: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In: Medical Image Computing and Computer Assisted Intervention, LNCS. vol. 15006, pp. 208–218. Springer (2024)
3. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the 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. Han, W., Yin, J., Jin, X., Dai, X., Shen, J.: BRNet: Exploring comprehensive features for monocular depth estimation. In: Proceedings of European Conference on Computer Vision. pp. 586–602 (2022)
6. Huang, B., Zheng, J.Q., Nguyen, A., Xu, C., Gkouzionis, I., Vyas, K., Tuch, D., Giannarou, S., Elson, D.S.: Self-supervised depth estimation in laparoscopic image using 3D geometric consistency. In: Medical Image Computing and Computer Assisted Intervention, LNCS. vol. 13437, pp. 13–22 (2022)
7. Hwang, M., Seita, D., Thananjeyan, B., Ichnowski, J., Paradis, S., Fer, D., Low, T., Goldberg, K.: Applying depth-sensing to automated surgical manipulation with a da Vinci robot. In: International Symposium on Medical Robotics. pp. 22–29 (2020)
8. Kingma, D.P., Ba, J.: ADAM: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Li, W., Hayashi, Y., Oda, M., Kitasaka, T., Misawa, K., Mori, K.: Spatially variant biases considered self-supervised depth estimation based on laparoscopic videos. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization pp. 1–9 (2021)
10. 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, LNCS. vol. 13434, pp. 467–477 (2022)
11. Li, W., Hayashi, Y., Oda, M., Kitasaka, T., Misawa, K., Mori, K.: Multi-view guidance for self-supervised monocular depth estimation on laparoscopic images via spatio-temporal correspondence. In: Medical Image Computing and Computer Assisted Intervention, LNCS. vol. 14228, pp. 429–439 (2023)
12. 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)
13. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., et al.: EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical Image Analysis **71**, 102058 (2021)
14. Qian, L., Zhang, X., Deguet, A., Kazanzides, P.: ARANIS: Augmented reality assistance for minimally invasive surgery using a head-mounted display. In: Medical Image Computing and Computer Assisted Intervention, LNCS. vol. 11768, pp. 74–82 (2019)
15. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J., Civera, J.: 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 (2021)
16. Sánchez-González, P., Cano, A.M., Oropesa, I., Sánchez-Margallo, F.M., Pozo, F.D., Lamata, P., Gómez, E.J.: Laparoscopic video analysis for training and image-guided surgery. Minimally Invasive Therapy & Allied Technologies **20**(6), 311–320 (2011)
17. Shao, S., Pei, Z., Chen, W., Wu, X., Li, Z.: NDDepth: Normal-distance assisted monocular depth estimation. In: Proceedings of the International Conference on Computer Vision. pp. 7931–7940 (2023)
18. Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B.: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. Medical Image Analysis **77**, 102338 (2022)
19. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 1164–1174 (2021)

20. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 10371–10381 (2024)
21. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2025)
22. Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R.: Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
23. Zhang, N., Nex, F., Vosselman, G., Kerle, N.: Lite-Mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 18537–18546 (2023)
24. Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S.: MonoViT: Self-supervised monocular depth estimation with a vision transformer. In: 2022 International Conference on 3D Vision. pp. 668–678 (2022)
25. Zhou, H., Greenwood, D., Taylor, S.: Self-supervised monocular depth estimation with internal feature fusion. In: British Machine Vision Conference (2021)
26. 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)