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

Wenda Li[1*], Yuichiro Hayashi[1], Masahiro Oda[2,1], Takayuki Kitasaka[3], Kazunari Misawa[4], Kensaku Mori[1,2,6]

1 Graduate School of Informatics, Nagoya University 2 Information Technology Center, Nagoya University 3 Faculty of Information Science, Aichi Institute of Technology 4 Aichi Cancer Center Hospital 5 Research Center of Medical Bigdata, National Institute of Informatics

INTRODUCTION

- **Depth Perception in Minimally Invasive Surgery (MIS)**
  - Narrow field of view (FoV) with poor depth visualization in MIS [1]
  - Two main solutions to view problem [2]
    - Robotic-assisted MIS ne[illegible] valu[illegible]ze automation
    - AR-assisted MIS need dep[illegible] for [illegible] reconstruction
  - → Depth value is important for the minimally invasive surgeries
- **Previous methods for monocular depth estimation**
  - A self-supervised learning strategy has been a mainstream method [3]
    - Leverage adjacent images to estimate relative poses of the camera
    - Calculate photometric error of matched pixels between adjacent images
  - Self-supervised depth estimation is introduced to laparoscopic scenes [3]

PURPOSE

- **Problems in current methods for laparoscopic images**
  - Smooth surface of organs and complex rotations of laparoscope lead depth estimation become more difficult
  - Previous method (GCDepthL [4]) enforced geometric constraints while overlooking local geometric structures

Purpose Enforce stronger geometric constraints for depth and pose estimation to predict accurate depths and 3D reconstruction

- **Contributions**
  - Propose a feature-matching process by calculating the 4D score volume
  - Introduce surface normal estimation and build the depth-normal consistency
  - Model an uncertainty map for the depth-normal consistency to alleviate bias

METHOD

- **Architecture of network with three branches**
  - Depth: input target image $\mathbf{I}_t$ (t means time t) and output depth map $\mathbf{D}_t$
  - Pose: input $\mathbf{I}_t$ and source images $\mathbf{I}_s$ (s means time $t-1$ or $t+1$) to predict relative pose $\mathbf{T}_{t\to s}$ (Fig. 1)
  - Normal vector: input $\mathbf{I}_t$ and $\mathbf{I}_s$ and output normal maps $\mathbf{N}_t$ and $\mathbf{N}_s$
- **Self-supervised learning strategy**
  - Extract feature maps $\mathbf{F}_t$ and $\mathbf{F}_s$ from $\mathbf{I}_t$ and $\mathbf{I}_s$ and construct a 4D score volume by feature-matching to predict poses
  - Match 2D locations p_t and p_s by predicted depth value $\mathbf{D}_t^{p_t}$ and generate synthesis images $\mathbf{I}_{s\to t}$, then calculate reprojection error as

$$\mathcal{L}_r = \frac{1}{|\mathrm{H}|}\sum_{p\in\mathrm{H}} \min \frac{\alpha}{2}\left(1-\mathrm{SSIM}(\mathbf{I}_t, \mathbf{I}_{s\to t}, p)\right) + (1-\alpha)\|\mathbf{I}_t^p - \mathbf{I}_{s\to t}^p\|_1$$

- **Depth-normal consistency under distance-based uncertainty**
  - As shown in Fig. 2, convert back-projected 3D positions $\mathbf{P}$ to normal map $\mathbf{N}_b$ by surrounding pixels' 2D locations p^i and p^j belonging to Ω as

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

  - Model uncertainty map $\mathbf{U}$ based on the distances d of points within local region π to the synthesized plane using as surrounding pixels' 2D locations p^k as

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

  - Construct consistency of converted normal and predicted normal maps as

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

  - Construct consistency of adjacent predicted normal maps as

$$\mathcal{L}_n = \frac{1}{|\mathrm{H}|}\sum_{p\in\mathrm{H}} (1-\mathbf{N}_{s\to t}^p\cdot\mathbf{N}_t^p)$$

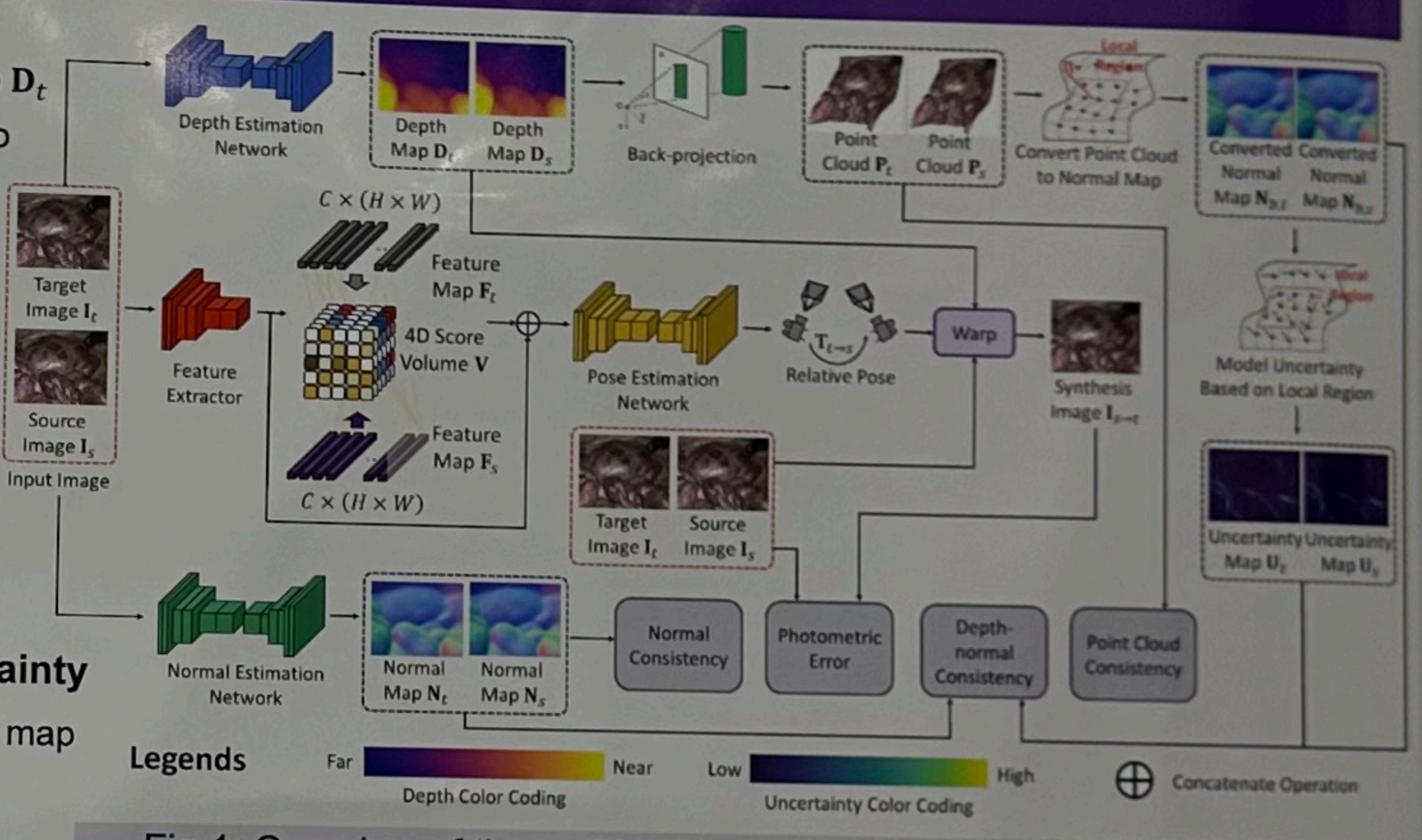

Fig 1. Overview of the architecture with optimized pose estimation

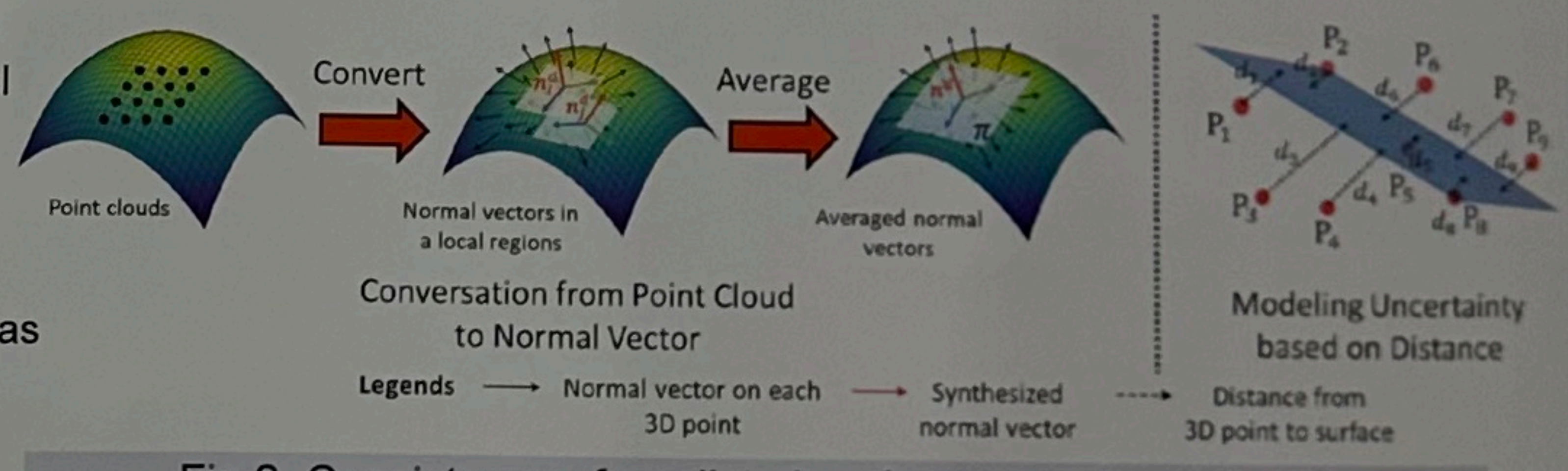

Fig 2. Consistency of predicted surface normal and depths

- Total loss function is combined with smoothness loss $\mathcal{L}_s$ and 3D points loss $\mathcal{L}_p$ [4] as

$$\mathcal{L}_f = \mathcal{L}_r + \lambda\mathcal{L}_c + \gamma\mathcal{L}_n + \mu\mathcal{L}_s + \xi\mathcal{L}_p$$

EXPERIMENTS & RESULTS

- **Datasets and evaluation metrics**
  - Datasets: laparoscopic datasets Hamlyn [5] and SCARED [6]
  - Training image size: downsampled to quarter size of original size
  - Evaluation metrics: 2D and 3D metrics for predicted depth maps
- **Implementation details**
  - In final total loss function, λ, γ, μ are set as 0.01and λ, ξ are 10^{-3}
  - ResNet18 with pretrained parameters is adopted as encoder
  - Decoder has 5 layers as the same as previous methods [3,4,7]

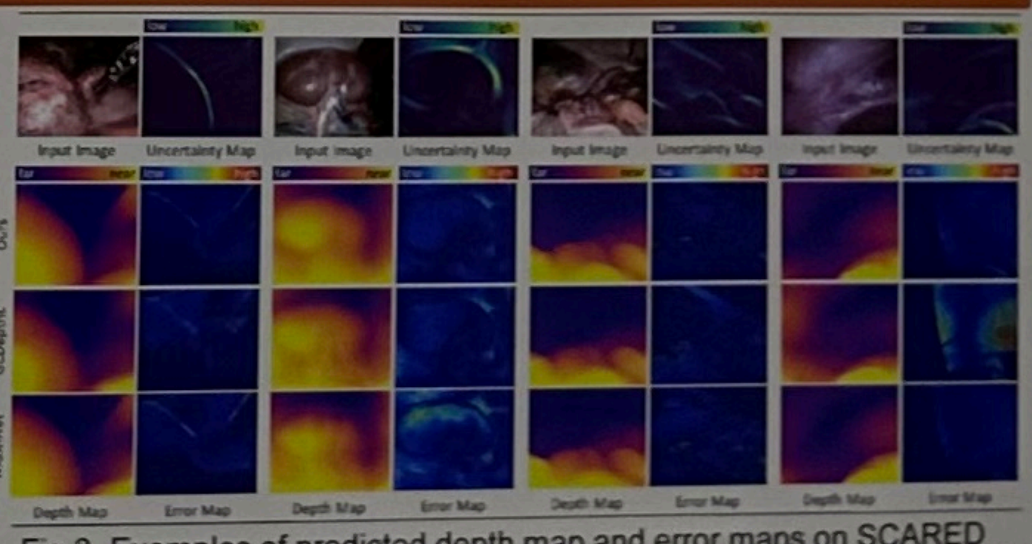

Fig 3. Examples of predicted depth map and error maps on SCARED

Table 1 Quantitative results on Hamlyn for depth estimation

Metrics		Ours	AF-SfMLearner [3]	GCDepthL [4]	MGMNet [7]
2D	Abs Rel ↓	0.143	0.169	0.162	0.159
	RMSE ↓	13.142	15.862	14.762	14.553
3D	Comp. ↓	5.202	6.577	5.881	6.114
	Recall ↑	0.437	0.342	0.399	0.369

Table 2 Quantitative results on SCARED datasets for depth estimation

Metrics		Ours	Ours w/o consistency	AF-SfMLearner [3]	GCDepthL [4]	MGMNet [7]
2D	Abs Rel ↓	0.055	0.061	0.066	0.062	0.063
	RMSE ↓	4.800	5.136	5.608	5.851	5.696
3D	Comp. ↓	2.435	2.840	3.234	2.625	2.798
	Recall ↑	0.777	0.738	0.703	0.729	0.717

DISCUSSION

- In Table 1 and Table 2, proposed method has better performance than previous methods on SCARED and Hamlyn datasets
- Proposed method outperformed on two 2D and two 3D metrics
- In Fig. 3, all existing methods output similar predicted depth maps, but proposed methods has lower errors through error maps
- Error maps reveal the error hidden error in the predicted depth maps even the depth maps perform smooth

CONCLUSION

- We construct a consistency of predicted depth maps and normal maps with an optimized pose estimation process via a novel 4D score volume
- Experiment results show our method has better performance on depth estimation

Ref.

[1] Geis, W.P.: Head-mounted video monitor for global visual access in mini-invasive surgery. Surgical Endoscopy 10(7), 768–770 (1996)
[2] Qian, L., et al.: Aramis: Augmented reality assistance for minimally invasive surgery using a head-mounted display. In: MICCAI 2019. LNCS, vol. 11768, pp. 74–82. Springer International Publishing, Cham (2019)
[3] Shao, S., et al: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. Medical image analysis 77:102338, 2022
[4] Li, W., et al.: Geometric constraints for self-supervised monocular depth estimation on laparoscopic images with dual-task consistency. In: MICCAI 2022: 25th International Conference, volume 13434, pp. 467–477, 2022
[5] Allan, M., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133, 2021
[6] Recasens, D, et al.: 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
[7] Li, W., et al.: Multi-view Guidance for Self-supervised Monocular Depth Estimation on Laparoscopic Images via Spatio-temporal Correspondence. In: MICCAI 2023: 26th International Conference, volume 14228, pp. 429–439, 2023