2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Learning-Based Depth and Pose Estimation for Monocular Endoscope with Loss Generalization

Aji Resindra Widya[1], Yusuke Monno[1], Masatoshi Okutomi[1],
Sho Suzuki[2], Takuji Gotoda[2], and Kenji Miki[3]

Abstract— **Gastroendoscopy has been a clinical standard for diagnosing and treating conditions that affect a part of a patient's digestive system, such as the stomach. Despite the fact that gastroendoscopy has a lot of advantages for patients, there exist some challenges for practitioners, such as the lack of 3D perception, including the depth and the endoscope pose information. Such challenges make navigating the endoscope and localizing any found lesion in a digestive tract difficult. To tackle these problems, deep learning-based approaches have been proposed to provide monocular gastroendoscopy with additional yet important depth and pose information. In this paper, we propose a novel supervised approach to train depth and pose estimation networks using consecutive endoscopy images to assist the endoscope navigation in the stomach. We firstly generate real depth and pose training data using our previously proposed whole stomach 3D reconstruction pipeline to avoid poor generalization ability between computer-generated (CG) models and real data for the stomach. In addition, we propose a novel generalized photometric loss function to avoid the complicated process of finding proper weights for balancing the depth and the pose loss terms, which is required for existing direct depth and pose supervision approaches. We then experimentally show that our proposed generalized loss performs better than existing direct supervision losses.**

I. Introduction

Gastroendoscopy is one of the golden standards for finding and treating abnormalities inside a patient's digestive tract, including the stomach. Even though gastroendoscopy gives enormous advantages for the patient, such as no need for invasive surgeries, it is known that there exist some challenges for medical practitioners, such as the loss of depth perception and the difficulty in assessing the endoscope pose. These challenges lead to difficulties in navigating and understanding the scene captured by the endoscope system, making the localization of a found lesion hard for the practitioners.

Previous studies have proposed to reconstruct the 3D model of a whole stomach with its texture [1]–[3] to provide a global view of the stomach and the estimated endoscope trajectory. It enables medical practitioners to perform a second inspection with more degree of freedom after an initial gastroendoscopy procedure. While the whole stomach 3D reconstruction provides the depth and the endoscope trajectory, the methods [1], [2] cannot be done alongside the gastroendoscopy procedure in real-time.

Recent developments in endoscopy systems introduce a stereo camera to provide real-time depth information [4]–[7]. While the stereo endoscope solves the lack of depth perception, a monocular endoscope is still the mainstream system in clinical practice. As an alternative to the stereo endoscope, deep learning-based approaches have been proposed to provide depth information for monocular endoscopy [8]–[10].

To effectively tackle the endoscope navigation and the lesion localization challenges, only providing depth information is not enough. Both continuous depth and pose information are needed to address these challenges appropriately. Both supervised and self-supervised deep-learning-based approaches are heavily adopted to address simultaneous depth and pose estimation [11]–[15]. A commonly used supervision approach is to take the direct Euclidean distance losses for the predicted depth and pose in comparison with the respective ground truths or references [11], [13]. In this approach, computer-generated (CG) and/or phantom models are commonly used for the training of depth and pose estimation networks, affecting the network generalization between CG and real data. In addition, the direct supervision approach needs balancing weights for depth and pose loss terms, which are difficult to search [16].

As a self-supervised approach, the study [12] uses consecutive frames as the inputs to train the network to simultaneously predict depth and pose by minimizing the photometric error of a view synthesis problem (image warping between consecutive frames based on the predicted depth and pose). Using the same principle, the study [14] trains a recurrent neural network to predict the depth and the pose and uses them as the inputs for standard SLAM [17] for further refinement. Since it needs an additional hand-crafted method bootstrapped to the network architecture, this approach is not trainable in an end-to-end manner. Even though a self-supervised training approach does not need labeled data for training and thus it is generally more convenient to train, its performance is yet to beat the supervised approach.

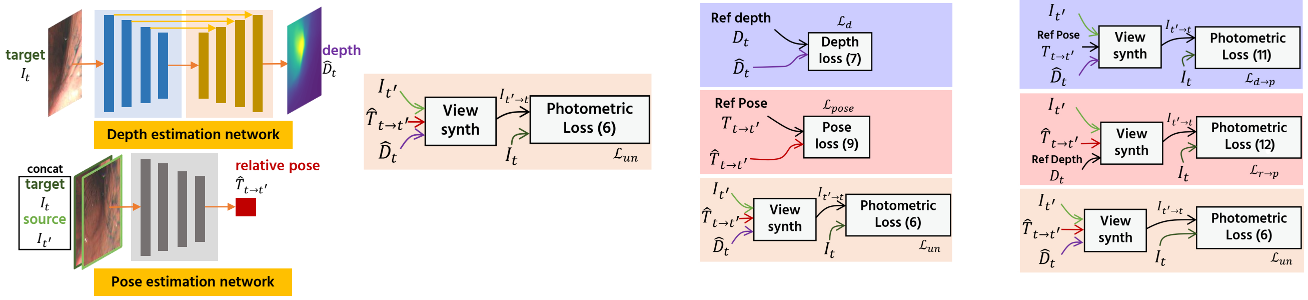
In this paper, we propose a supervised approach to simultaneously train depth and pose networks using consecutive images for monocular endoscopy of the stomach. To avoid the generalization problem between CG and real data, we

This work was supported by JSPS KAKENHI Grant Number 17H00744 and 21K12737.

[1] A. R. Widya, Y. Monno, and M. Okutomi are with the Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan (e-mail: aresindra@ok.sc.e.titech.ac.jp; ymonno@ok.sc.e.titech.ac.jp; mxo@sc.e.titech.ac.jp).

[2] S. Suzuki and T. Gotoda are with the Division of Gastroenterology and Hepatology, Department of Medicine, Nihon University School of Medicine, Chiyoda-ku, Tokyo 101-8309, Japan.

[3] K. Miki is with the Department of Internal Medicine, Tsujinaka Hospital Kashiwanoha, Kashiwa-city, Chiba 277-0871, Japan.



(a) Network structure [18].

(b) Self-supervised photometric loss [18].

(c) Direct depth and pose supervision [10], [11] + photometric loss.

(d) Proposed generalized photometric loss.

Fig. 1: The network structure which consists of depth and pose estimation networks is shown in (a). Figures (b)-(d) show the comparison between the existing self-supervised photometric loss, the existing direct depth and pose supervision loss, and our proposed generalized photometric loss. In both (c) and (d), the loss in the purple-colored box is used for training the depth estimation network and the loss in the pink-colored box is used for training the pose estimation network. The existing depth and pose supervision approach trains the depth and the pose estimation networks by directly taking the Euclidean distance between the predicted depth and its reference and also between the predicted pose and its reference, respectively. This direct supervision approach needs balancing the weights for each loss term, which are difficult to search, because their physical meanings are different. In our proposed generalize loss, we adjusted our loss terms so that each of them has the same physical meaning, i.e., the photometric error. This generalization eliminates the need for the balancing weight search.

apply the whole stomach 3D reconstruction pipeline [2] to generate reference depth and pose from real endoscope data for network training. Additionally, we propose a novel loss generalization by unifying the depth and the pose losses into a photometric error loss for our supervised training to avoid the necessity of delicate weight balancing between the depth and the pose losses. Finally, we show that our supervised training with a novel generalized loss function has better performance than the existing direct depth and pose supervision. Our method achieves up to 60fps at test time for depth and pose predictions.

II. Materials and Methods

Figure 1(a) overviews the network structures used in our experiment and Figure 1(b)-(d) show how we train them using three different methods, i.e., an existing self-supervised photometric loss [18], an existing direct depth and pose losses [10], [11] combined with the photometric loss, and our proposed generalized photometric loss. In this section, we firstly explain the endoscope dataset (Section II-A). We then review the existing self-supervision and direct supervision methods (Section II-B, II-C). Finally, we explain our proposed loss generalization (Section II-D).

A. Training data generation

In this work, we used the same endoscope video dataset from our previous work [2]. We used six subjects' endoscope videos undergone general gastroendoscopy procedure. We then extracted all the image frames from all the videos and used them as training and testing data. The experimental protocol was approved by the research ethics committee of Tokyo Institute of Technology and Nihon University Hospital.

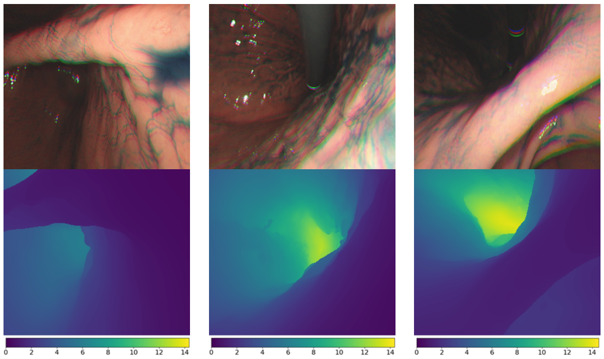


Fig. 2: Some examples of the generated reference depth based on the estimated camera poses and the obtained whole stomach 3D model using [2]. We can see that the generated reference depth images reflect the structures seen in the color images.

For the direct supervision (Section II-C) and the proposed loss generalization (Section II-D) methods, we used the previously extracted real endoscope image frames to generate the reference depth and pose. For this purpose, we firstly applied the whole stomach 3D reconstruction pipeline [2] and then extracted the reference depth and pose from the generated whole stomach 3D model and the estimated endoscope poses. Figure 2 shows some of the RGB images and the generated reference depths. We used the generated reference depth and pose for both training and testing.

B. Self-supervised depth and pose estimation

Our depth and pose estimation networks are inspired by monodepth2 architecture [18] shown in Figure 1(a). It consists of two separate networks, each for depth and pose estimation purpose. Both networks are trained together to learn a view-synthesis problem, i.e., to predict the appearance of a target image given a view point of another image by minimizing its photometric error.

Let I_t be a target frame and $I_{t'}$ be a source frame. The objective of the network is to minimize a photometric error *pe* between a target frame and a warped source frame. In general, a photometric error *pe* between two images A and B can be defined using pixel value difference (L1) and structural similarity index measure (SSIM) [19] such that

$$pe(A,B) = \frac{\alpha}{2}(1 - \text{SSIM}(A,B)) + (1-\alpha)\|A-B\|_1 \quad (1)$$

where α is the balancing term between the L1 and the SSIM terms. Let $\hat{D}_t$ be the predicted depth of the target frame I_t, $\hat{T}_{t\to t'}$ be the predicted relative pose from the target frame I_t to the source frame $I_{t'}$, and K be a calibrated camera intrinsic parameters. We then define the view synthesis (image warping) problem as

$$I_{t'\to t} = \epsilon(I_{t'}, \pi(\hat{D}_t, \hat{T}_{t\to t'}, K)) \quad (2)$$

where π is a function to project the pixel coordinate of target image I_t in source image $I_{t'}$ and ϵ is a pixel sampling function based on the projected pixel coordinate given by π. In our implementation, we used two consecutive frames as our source frames, *i.e.*, I_{t+1} and I_{t-1}. Instead of averaging the photometric error *pe* for each warped source frame, we simply take the minimum such that the final pixel-wise photometric error can be expressed as

$$\mathcal{L}_p = \min_{t'\in(t+1,t-1)} pe(I_t, I_{t'\to t}) \quad (3)$$

To ensure that only the reliable pixels are optimized, we masked out the non-reliable pixel using the automask [18] defined by a logical operation as

$$\mu = [\min_{t'\in(t+1,t-1)} pe(I_t, I_{t'\to t}) < \min_{t'\in(t+1,t-1)} pe(I_t, I_{t'})] \quad (4)$$

We also used edge-aware smoothness so that there is no discontinuities in the predicted depth [20] that can be expressed as

$$\mathcal{L}_s = |\partial_x \hat{d}^*_t| e^{-|\partial_x I_t|} + |\partial_y \hat{d}^*_t| e^{-|\partial_y I_t|} \quad (5)$$

where ∂_x and ∂_y is the partial derivative on each x and y direction and $\hat{d}^*_t$ is the mean-normalized predicted inverse depth [21]. The final self-supervised loss function consists of the masked photometric error and the smoothness term as

$$\mathcal{L}_{un} = \frac{1}{N}\sum_i^N \mu^i \mathcal{L}^i_p + \lambda \mathcal{L}^i_s \quad (6)$$

where i represents a pixel index, N is the total number of the pixels, and λ is the balancing weight between the photometric error and the depth smoothness loss.

C. Supervised depth and pose estimation

To supervise the depth estimation network, we followed the method [22] and used inverse depth d instead of depth D. We followed a standard inverse depth error loss function which compares the inverse depth prediction $\hat{d}_t$ and its reference inverse depth d_t. It consists of three sub-components which can be formulated as follows

$$\mathcal{L}_d = |\hat{d}_t - d_t|, \quad (7a)$$

$$\mathcal{L}_g = |\partial_x(\hat{d}_t, d_t)| + |\partial_y(\hat{d}_t, d_t)|, \quad (7b)$$

$$\mathcal{L}_{\text{SSIM}} = \frac{1 - \text{SSIM}(\hat{d}_t, d)}{2} \quad (7c)$$

The total loss for depth supervision can finally be expressed as

$$\mathcal{L}_{d_t} = \frac{1}{N}\sum_i^N 0.1\mathcal{L}^i_d + \mathcal{L}^i_g + \mathcal{L}^i_{\text{SSIM}}. \quad (8)$$

For the pose estimation supervision, the common practice is to directly supervise the pose estimation network by measuring Euclidean distance between the predicted relative pose and its reference pose [11], [16], [23], i.e.,

$$\mathcal{L}^{t\to t'}_{pose} = \zeta\|\hat{\mathbf{x}}_{t\to t'} - \mathbf{x}_{t\to t'}\|_2 + \theta\|\hat{\mathbf{r}}_{t\to t'} - \mathbf{r}_{t\to t'}\|_2 \quad (9)$$

where $\mathbf{x}$ is the translation vector component and $\mathbf{r}$ is the rotation vector components in the axis-angle representation from the relative pose $T_{t\to t'}$. The translation and rotation terms are balanced by ζ and θ as weights. To tie $\mathcal{L}_d$ and $\mathcal{L}_{pose}$ together, a photometric loss $\mathcal{L}_p$ is added to the direct supervision loss. Finally, the total supervised loss can be expressed as

$$\mathcal{L}_{su} = \frac{1}{N}\sum_i^N(\psi\mu^i\mathcal{L}^i_p + \gamma\mathcal{L}^i_{d_t}) + \sum_{j\in t'}\mathcal{L}^{t\to j}_{pose} \quad (10)$$

where ψ and γ are balancing weights for depth and pose losses.

D. Loss generalization

Since each of the components in (10) in the commonly used supervised loss has different physical meaning, the weight of each component has to be carefully selected, which is very difficult and usually performed in an empirical manner. It is also common that different weight balancing is needed for different kinds of environment such as outdoor and indoor scenes [16]. To address this limitation, we propose a novel depth and pose supervision loss function by generalizing the depth and the pose errors into the same physical meaning, i.e, a photometric error.

As illustrated in Figure 1(d), in order to generalize the loss into a photometric error, we use the reference relative pose $T_{t\to t'}$ to train the depth estimation network by optimizing the predicted depth $\hat{D}_t$ such that

$$\mathcal{L}_{d\to p} = \min_{t'\in(t+1,t-1)} pe(I_t, \epsilon(I_{t'}, \pi(\hat{D}_t, T_{t\to t'}, K))) \quad (11)$$

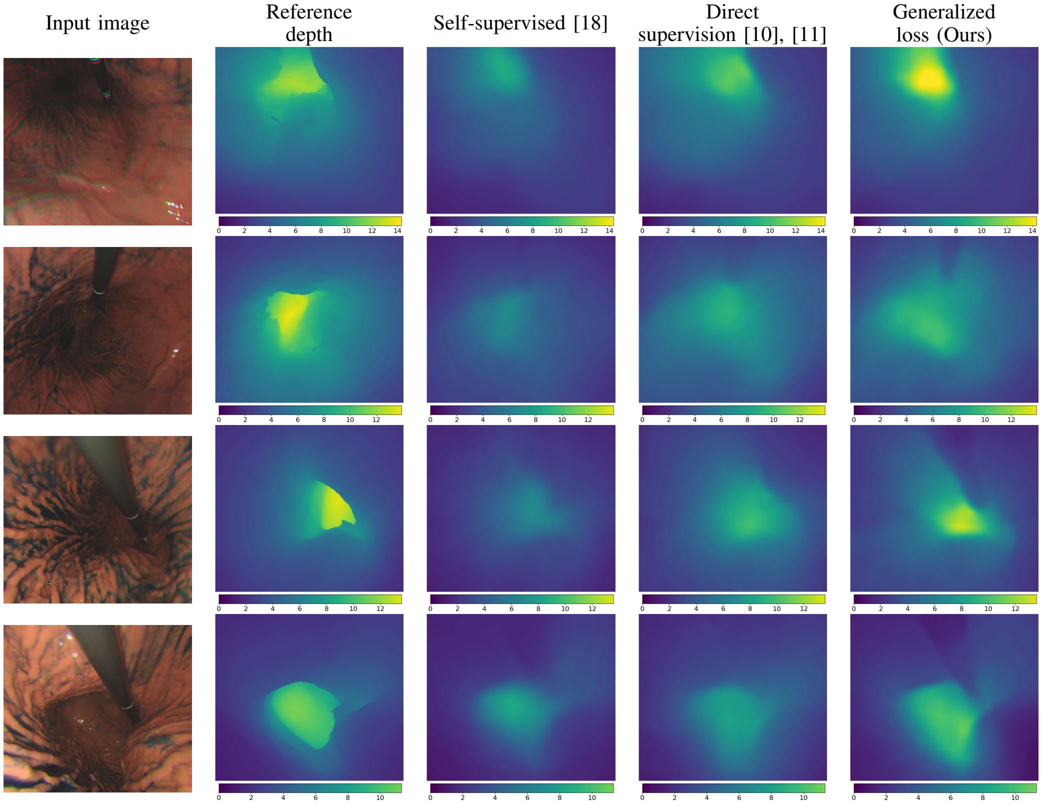


Fig. 3: Some examples of depth estimation results. Here we show the RGB images for better visualization, though we actually used red channel images as the input of the network according to the finding in [2]. We compare the depth prediction results of the self-supervision [18], the direct supervision [11], [16], and our proposed generalized loss supervision. As we can see, our proposed method not only estimates closer depth to the reference, but also better estimates the structures and the boundaries, including the endoscope rod. In some cases, the direct supervision results are too smooth.

Conversely, we use the reference depth D_t to train the pose estimation network by optimizing the predicted relative pose $\hat{T}_{t\to t'}$ such that

$$\mathcal{L}_{r\to p} = \min_{t'\in(t+1,t-1)} pe(I_t, \epsilon(I_{t'}, \pi(D_t, \hat{T}_{t\to t'}, K))) \tag{12}$$

The term described in (11) can be defined as *depth loss w.r.t reference pose as photometric loss* while the term described in (12) as *pose loss w.r.t reference depth as photometric loss*. We also calculate the reliable pixel masks, $\mu_{d\to p}$ and $\mu_{r\to p}$, for each $\mathcal{L}_{d\to p}$ and $\mathcal{L}_{r\to p}$ respectively using the same principle as (4).

Combining (11) and (12) with (3) to tie the depth and the pose network training together, we can write the final loss function as

$$\begin{aligned}\mathcal{L}_{gen} = \frac{1}{N}\sum_{i}^{N}(\mu^i_{d\to p}\mathcal{L}^i_{d\to p} + \mu^i_{r\to p}\mathcal{L}^i_{r\to p} \\ + \underbrace{\mu^i\mathcal{L}^i_p + \lambda\mathcal{L}^i_s)}_{\mathcal{L}_{un}\ (6)}\end{aligned} \tag{13}$$

which eliminates the intricate search of balancing weights for the depth and the pose loss terms.

III. Results and Discussion

A. Implementation details

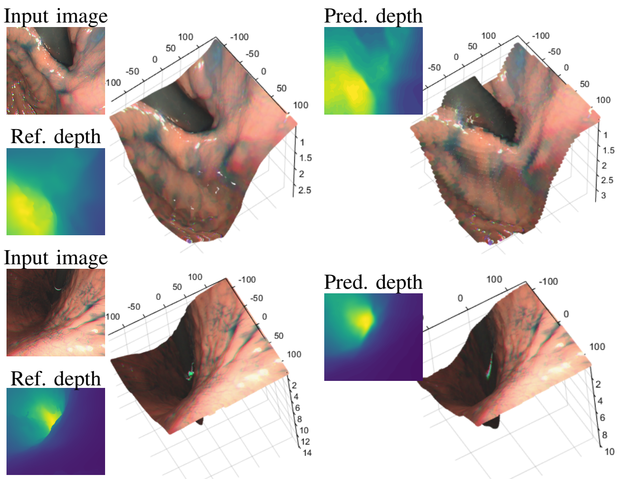
Following [18], we used ResNet-18 architecture [24] for our depth and pose estimation networks. We simultaneously trained our depth and pose estimation networks using a single NVIDIA GeForce GTX 1080Ti GPU. Our networks were trained for 100 epochs with the learning rate of 10^{-4} with the decay factor of 10^{-1} after 50 epochs. We set the term weights for the self-supervised and the generalized loss training as $\alpha = 0.85$ and $\lambda = 0.001$. Additionally, we set the extra balancing weights for the direct supervision as $\gamma = 30$, $\zeta = \psi = 15$, and $\theta = 160$.

We divided six subjects into four subjects for training (Subjects 3-6, 9000 images) and two subject for testing (Subjects 1-2, 2350 images). The image resolution is 288×256 pixels. Following the finding of our previous research to tackle the color channel misalignment problem [2], we used only red channel images to train the networks.

B. Depth estimation evaluation

Figure 3 shows the subjective evaluation results. We evaluated the relative depth error and the depth accuracy as

- Relative error: $\frac{|\hat{D}_t - D_t|}{D_t}$
- Depth accuracy: $\delta = \max(\frac{D_t}{\hat{D}_t}, \frac{\hat{D}_t}{D_t})$



(a) Reference depth (b) Generalized loss (Ours)

Fig. 4: Comparison of the generated 3D point clouds using the reference depth and the predicted depth by our proposed method. As we can see, both the depth and the structure from the predicted depth are close to the ones generated from the reference depth.

TABLE I: Depth estimation objective evaluation.

Method	Accuracy ↑			Relative errors ↓		
	$\delta < 1.25^1$	$\delta < 1.25^2$	$\delta < 1.25^3$	mean	max	median
Test on Subject 1						
Self-supervised [18]	0.374	0.703	0.838	0.635	8.812	0.286
Direct supervision [10], [11]	0.525	**0.814**	0.900	0.432	6.628	**0.209**
Generalized loss (Ours)	**0.540**	0.804	**0.902**	**0.416**	**5.445**	0.212
Test on Subject 2						
Self-supervised [18]	0.477	0.767	0.867	0.472	8.673	0.227
Direct supervision [10], [11]	0.536	0.806	0.910	0.349	6.717	**0.210**
Generalized loss (Ours)	**0.579**	**0.822**	**0.916**	**0.336**	**6.632**	0.213
Test on Training data						
Self-supervised [18]	0.565	0.819	0.912	0.342	3.602	0.204
Direct supervision [10], [11]	**0.961**	**0.992**	**0.996**	**0.064**	**0.465**	**0.059**
Generalized loss (Ours)	0.791	0.916	0.956	0.168	1.313	0.114

*Numbers are dimensionless

where D_t is the reference depth and $\hat{D}_t$ is the predicted depth. For the relative error, we calculated the errors for all the pixels of all the frames and evaluated the mean, the maximum, and the median values. For the depth accuracy evaluation, we measured the ratio between the number of pixels that have a lower error than a threshold controlled by k (i.e., $\delta < 1.25^k$) and the total number of the pixels.

Table I shows the objective evaluation results. Since the predicted depth is only up to scale, we scaled the predicted depth by minimizing the average RMSE for the entire sequence. From Table I, we can see that our proposed method has better performance compared to the self-supervised method by a fair margin. In addition, our proposed method generally shows better performance compared to the direct supervision method. Even though our proposed method comes seconds in the median relative error, the values are very close. In addition to the testing on the test data (Subject 1 and 2), we also tested each of the trained networks on the training data. As we can see, the direct supervision has the best results for this evaluation. However, it can be noticed that the performance of the direct supervision on the test data falls sharply compared to its performance on the training data. It shows that the depth estimation network trained with the direct supervision has poor generalization to the data that have never been seen during the training.

TABLE II: Pose estimation objective evaluation.

Method	Rotation error ↓			Translation error ↓		
	mean	max	median	mean	max	median
Test on Subject 1						
Self-supervised [18]	0.562	0.809	0.519	0.253	0.531	0.211
Direct supervision [10], [11]	0.579	0.887	0.539	0.274	0.555	0.226
Generalized loss (Ours)	**0.458**	**0.714**	**0.426**	**0.222**	**0.471**	**0.178**
Test on Subject 2						
Self-supervised [18]	0.581	0.802	0.564	0.283	0.587	0.239
Direct supervision [10], [11]	0.606	0.836	0.588	0.296	0.578	0.243
Generalized loss (Ours)	**0.517**	**0.742**	**0.491**	**0.246**	**0.495**	**0.206**
Test on Training data						
Self-supervised [18]	0.554	0.769	0.511	0.276	0.585	0.231
Direct supervision [10], [11]	**0.195**	**0.324**	**0.172**	**0.116**	**0.265**	**0.093**
Generalized loss (Ours)	0.385	0.544	0.355	0.182	0.371	0.154

*Numbers are dimensionless

Figure 4 shows the resulting color point clouds by fusing a single image with its predicted depth. As we can see, the resulting point cloud from our predicted depth is very close to the point cloud from the reference depth.

C. Pose estimation evaluation

For pose estimation evaluation, we first split the full sequences of Subject 1 and 2 into the groups of 150 consecutive frames. The predicted poses were then aligned with the reference poses using Umeyama transform [25]. We then used absolute pose error (APE) to evaluate the translation and rotation components of the predicted poses $\hat{P} \in \mathrm{SE}(3)$ against the reference poses $P \in \mathrm{SE}(3)$. Given the absolute relative pose between a pair of predicted pose and its ground truth $E = P^{-1}\hat{P}$, the translation and the rotation errors can be defined as

$$APE_{rot} = \|\mathrm{rot}(E) - I_{3\times3}\|_F \tag{14}$$

$$APE_{trans} = \|\mathrm{trans}(E)\|_2 \tag{15}$$

where $\|.\|_F$ is Frobenius norm. We then averaged all the obtained APEs over all of the evaluation points. The objective evaluation results can be seen in Table II.

From Table II, we can see that, based on the evaluation on the test data, our proposed generalized loss has better performance compared to the self-supervised and the direct supervision methods. Even though it is evident that the direct supervision has the best result when tested on the training data, its performance drops sharply when tested on the test data. This characteristic is consistent with the results previously shown in the depth estimation evaluation. This is because that the direct supervision losses induce poor generalization performance. In addition, even without the intricate search of the term-balancing weights, our method could achieve the best result for the test data.

Finally, we show the trajectory prediction results in Figure 5, including the trajectory prediction result from ORB-

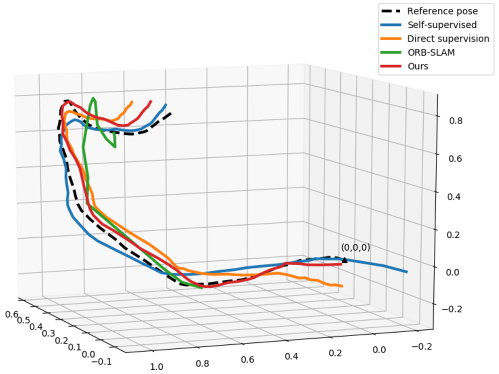


Fig. 5: The predicted trajectory for a sample sequence. As we can see, our prediction result is the closest to the reference. ORB-SLAM [26] could only predict the poses of 16 frames among 150 input frames.

SLAM [26]. As we can see, our prediction result is the closest to the reference, while ORB-SLAM could only predict the poses of 16 frames among the 150 input frames.

IV. Conclusions

In this paper, we have presented a novel generalized photometric loss for learning-based depth and pose estimation with monocular endoscopy. Compared to commonly used direct depth and pose supervision losses, which have different physical meanings for each loss term, we have proposed the generalized loss so that each of the loss terms has the same physical meaning, which is a photometric error. We have experimentally shown that our generalized loss supervision performs better than the direct depth and pose supervision without the need for an intricate search of term-balancing weights. We have also found that the generalization performance from train to test data of our proposed method is better than that of the direct supervision. In future work, we plan to fuse both depth and pose predictions from multiple frames for real-time 3D reconstruction and use the light source information to improve the depth and pose estimation. Additional related results be accessed from (http://www.ok.sc.e.titech.ac.jp/res/Stomach3D/).

References

[1] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Stomach 3D reconstruction using virtual chromoendoscopic images," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, no. 1700211, pp. 1–11, 2021.

[2] ——, "Whole stomach 3D reconstruction and frame localization from monocular endoscope video," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, no. 3300310, pp. 1–10, 2019.

[3] ——, "Stomach 3d reconstruction using virtual chromoendoscopic images," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–11, 2021.

[4] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov, "Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery," *Medical Image Analysis*, vol. 17, no. 8, pp. 974–996, 2013.

[5] J. Geng and J. Xie, "Review of 3-D endoscopic surface imaging techniques," *IEEE Sensors Journal*, vol. 14, no. 4, pp. 945–960, 2014.

[6] S.-P. Yang, J.-J. Kim, K.-W. Jang, W.-K. Song, and K.-H. Jeong, "Compact stereo endoscopic camera using microprism arrays," *Optics Letters*, vol. 41, no. 6, pp. 1285–1288, 2016.

[7] N. Mahmoud, C. Toby, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE Trans. on Medical Imaging*, vol. 38, no. 1, pp. 79–88, 2019.

[8] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *Int. Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1167–1176, 2019.

[9] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Medical Image Analysis*, vol. 48, pp. 230–243, 2018.

[10] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Self-supervised monocular depth estimation in gastroendoscopy using GAN-augmented images," in *Proc. of SPIE (Medical Imaging: Image Processing)*, vol. 11596, pp. 1 159 616–1–10, 2021.

[11] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots," *Neurocomputing*, vol. 275, pp. 1861–1870, 2018.

[12] M. Turan, E. P. Ornek, N. Ibrahimli, C. Giracoglu, Y. Almalioglu, M. F. Yanik, and M. Sitti, "Unsupervised odometry and depth learning for endoscopic capsule robots," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1801–1807, 2018.

[13] R. J. Chen, T. L. Bobrow, T. Athey, F. Mahmood, and N. J. Durr, "SLAM endoscopy enhanced by adversarial depth prediction," in *Proc. of KDD Workshop on Applied Data Science for Healthcare*, 2019.

[14] R. Ma, R. Wang, S. Pizer, J. Rosenman, S. K. McGill, and J.-M. Frahm, "Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions," in *Proc. of Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 573–582, 2019.

[15] K. B. Ozyoruk, G. I. Gokceler, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, H. Sahin, H. Araujo, H. Alexandrino, N. J. Durr, H. B. Gilbert, and M. Turan, "EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Medical Image Analysis*, 2021.

[16] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DoF camera relocalization," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2938–2946, 2015.

[17] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[18] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 3828–3838, 2019.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[20] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 270–279, 2017.

[21] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2022–2030, 2018.

[22] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.

[23] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5038–5047, 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[25] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Computer Architecture Letters*, vol. 13, no. 4, pp. 376–380, 1991.

[26] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.