Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Scale-Preserving Shape Reconstruction from Monocular Endoscope Image Sequences by Supervised Depth Learning

Takeshi Masuda[1] | **Ryusuke Sagawa**[1] | **Ryo Furukawa**[2] | **Hiroshi Kawasaki**[3]

[1] Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan

[2] Faculty of Engineering, Kindai University, Hiroshima, Japan

[3] Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan

Correspondence
Corresponding author: Takeshi Masuda,
Email: t.masuda@aist.go.jp

Present address
Central 1, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8560 Japan.

Abstract

Reconstructing 3D shapes from images are becoming popular, but such methods usually estimate relative depth maps with ambiguous scales. We propose a method for reconstructing a scale-preserving 3D shape from monocular endoscope image sequences through training an absolute depth prediction network. We first created a dataset of synchronized sequences of RGB images and depth maps using an endoscope simulator. Then, we train a supervised depth prediction network that estimates a depth map from a RGB image minimizing the loss compared to the ground-truth depth map. The predicted depth map sequence is aligned to reconstruct a 3D shape. Finally, the proposed method is applied to a real endoscope image sequence.

KEYWORDS
endoscope, monocular depth estimation, unsupervised depth estimation, SLAM, endoscope simulator.

1 | INTRODUCTION

Endoscopes are devices that allow non-invasive, direct observation of internal body structures and are used in the medical and industrial fields. In gastrointestinal examinations and surgeries, wired endoscopes are mainly used for the digestive tract, excluding the small intestine, while wireless capsule endoscopes are used for the examination of the entire digestive tract.

Most endoscopes are equipped with a monocular camera. Stereoscopic endoscopes are used to provide a 3D view to the surgeon in medical surgery, although they are not yet widely used due to operability limitations. Attempts are being made to reconstruct 3D information from stereoscopic endoscopes[1].

Reconstructing 3D structures from images has been an important topic in the history of computer vision. Recently, monocular 3D reconstruction has been realized using neural networks. Eigen et al.[2] developed a supervised monocular depth estimation using CNNs, whose depth error measure is scale-invariant, so the estimated depth is relative. Monocular 3D reconstruction is possible in limited domains where training images are provided, such as street views in the KITTI dataset[3] and indoor scenes in the NYU dataset[4]. Most methods for obtaining 3D structure from monocular images provide relative depth estimation.

Methods such as Structure from Motion (SfM) and Simultaneous Localization And Mapping (SLAM) are re-modeled by the learning framework[5] to reconstruct object shape and camera motion from image sequences. SfMLearner[6] modeled SLAM with two neural networks: DispNet for estimating disparity from a single image and PoseNet for estimating camera motion in a short-term image sequence. These two networks are jointly optimized by minimizing photometric consistency loss, which evaluates the change in pixel values before and after camera motion. Wang et al.[7] extended PoseNet with differentiable Direct Visual Odometry (DVO), and Godard et al.[8] (monodepth2) integrated PoseNet with ResNet-18 and introduced a binary mask to remove irregular motion. These methods do not require the ground truth depth map for training and are referred to as unsupervised methods.

Abbreviations: SfM, Structure from Motion; SLAM, Simultaneous Localization And Mapping; CNN, Convolutional Neural Network; MAE, Mean Absolute Error; ICP: Iterative Closest Point.

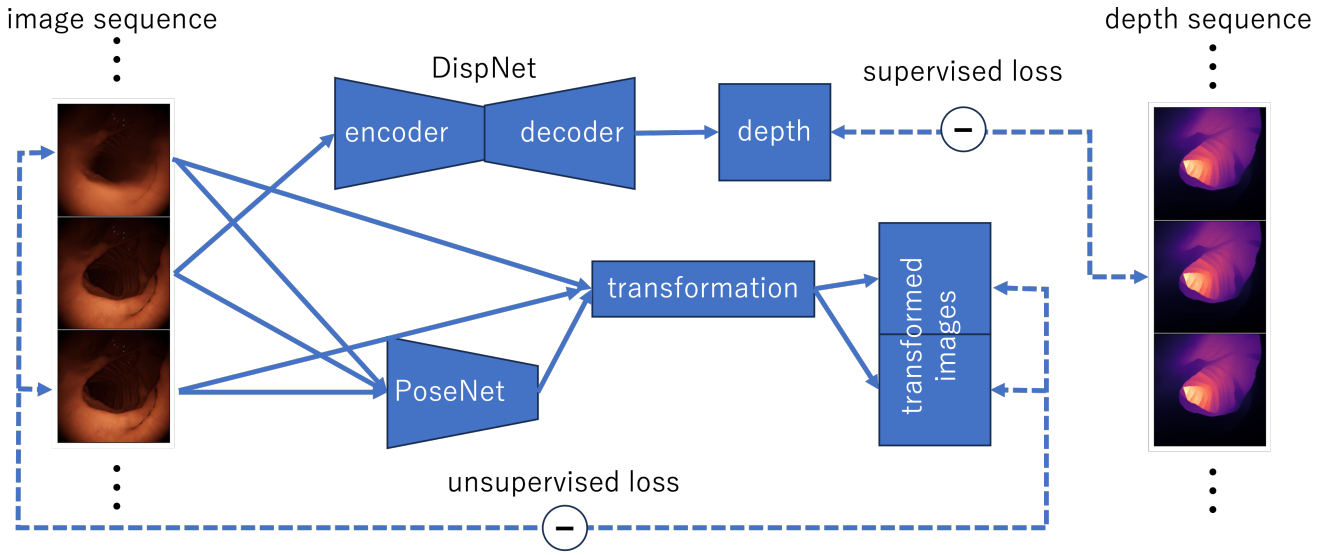

FIGURE 1 Diagram of the method. We assume that we have synchronized sequences of RGB images and depth maps. Is is composed of DispNet and PoseNet, and they are evaluated by the supervised and unsupervised losses.

Even if the depth has scale ambiguity, it can be constrained to be consistent throughout the image sequence. In[9], geometry consistency loss was introduced to penalize depth inconsistency between consecutive frames. This work was further improved by pre-correcting for rotation in[10] and by combining it with an additional relative monocular depth estimation module for high-density reconstruction that allows deformations in[11].

SfMLearner[6] was adapted to endoscopic image sequences: it is EndoSfMLearner[12]. They introduced geometry consistency loss, which measures depth consistency in addition to RGB values as well as photometric consistency loss. In this study, a capsule endoscope simulator called VR-Caps[13] was used to train the dataset by creating its own dataset. The obtained depth maps still have some scale ambiguity and they need to be rescaled appropriately for evaluation.

Correct scale is important for real medical applications. If a reference object is available, scale can be recovered by post-processing, but it would be beneficial if absolute shape and position could be predicted online without a reference object. DispNet[6] can be thought of as an encoder-decoder model, and the encoder part can be replaced by various image encoder networks. Fang et al.[14] tested various encoders for supervised and unsupervised learning of DispNet. Their conclusion is that supervised depth estimation is superior to unsupervised depth estimation.

This paper presents an absolute depth estimation method for monocular endoscopic image sequences. In general, monocular depth reconstruction is possible by assuming a specific scene. Monocular depth estimation is possible for endoscopic images of the digestive organs because the organs are often cylindrical in shape and the color range is limited, although there are individual differences. Compared to typical 3D reconstructed scenes such as KITTI and NYU, the difficulties in analyzing endoscopic scenes are featurelessness, wet specular reflections, uneven illumination, and severe turbulence and deformation. Therefore, using VR-Caps[13], we created a unique dataset that simulates a gastrointestinal endoscopy sequence with a synchronized sequence of RGB image and depth map pairs. Using this dataset, we first train DispNet in a supervised manner using truthful depth maps that preserve absolute depth, and then reconstruct the 3D shape by aligning and integrating the predicted depth maps. Finally, the developed method was applied to a real endoscopic image sequence.

The contributions of this paper are: first, we generated a synthetic dataset of the sequences of images with synchronized depth maps by an endoscope simulator, then we used it for training the DispNet to predict a scale preserving depth map from each image, and finally, we developed a SLAM algorithm to reconstruct the 3D shape from the predicted depth and applied it to a real endoscope image sequence.

We first explain dataset preparation in Sec. 2, then the supervised training of DispNet in Sec. 3 and PoseNet in Sec. 4. In Sec. 5, a SLAM method to reconstruct the 3D shape is proposed, and we show the results of application to the real endoscope image sequences in Sec. 6.

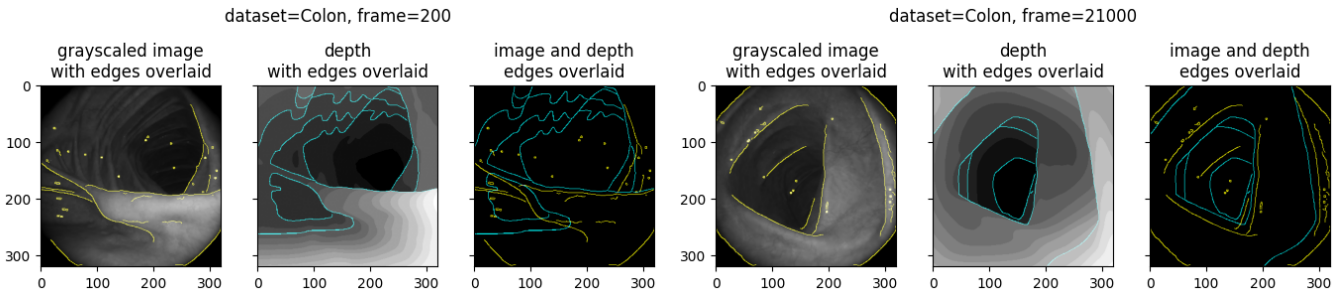

FIGURE 2 Two examples of misalignment of image and depth map in the VR-Caps dataset[13]. In each example, Left and Middle show the RGB image and the depth map at the same frame in gray-scale. For clear visualization, we extracted the image edges by Canny operator, and overlaid them on the gray scaled image and the depth map in yellow and cyan respectively. These two-colored edges are overlaid in Right, which shows that the edges of image and depth map are not perfectly aligned. The amount and orientation of misalignment are not consistent over the whole sequence.

2 | DATASET PREPARATION

For supervised training, ground truth depth maps are needed in addition to the RGB images. Currently, it is still difficult to obtain reliable real depth maps synchronized with the image frames from endoscopes. Instead, we used VR-Caps[13], a Unity[15]-powered capsule endoscopy simulator, to generate computer-graphics datasets. Its graphical shape model is derived from a human CT scan and includes the major digestive organs from the stomach to the large intestine, with realistic textures applied. It features configurable camera, reflection, and illumination parameters, and can synchronously render a pair of realistic RGB images and depth maps by controlling camera position and pose.

In the original VR-Caps dataset[16], the image size is 320×320 and the projection matrix and lens distortion are set to simulate a real capsule endoscope. In their dataset, we found that the synchronization between the RGB images and the depth map is incomplete (Fig. 2). We suppose that this misalignment is caused by on-line capturing, and we recreated our own dataset using the same graphics model by rendering the RGB images and the depth map off-line with perfect synchronization. For generating our own dataset, the image size is kept the same as 320×320 for both the RGB images and depth maps, and the focal length is 159.45 pixels and the projection center is at the image center. We didn't apply lens distortion because it can be easily undistorted even for real image cases.

With these settings, we manually controlled the camera position and pose, and finally generated 17 sequences of 9714 frames in total, and each frame is composed of synchronized frames of a RGB-image and a depth map. We split this dataset into two subsets: 13 sequences of 8249 frames for training and 4 sequences of 1465 frames for validation.

3 | SUPERVISED DEPTH ESTIMATION

3.1 | DispNet

DispNet is a neural network model that estimate a disparity map from a single RGB image (Fig. 1). It consists of an encoder and a decoder, where the former extracts image features and the latter synthesizes the disparity map. The supervised DispNet[14] was trained on the ground truth depth maps and RGB images with validate various encoder structures. Following many SfMLearner variants, the depth map D is obtained by taking the inverse of the disparity map d as $D = 1/d$.

The decoder part of DispNet consists of four-layers of up-convolution blocks. Each output of the up-convolution block is followed by a sigmoid function and a linear transformation, where $d' = \alpha\,\mathrm{sigmoid}(d) + \beta$ where d and d' are the disparity map before and after the transformation. The parameters α and β are fixed to match the output range of the sigmoid function $[-1, 1]$ to the required disparity range according to the target dataset, which is common to many successors of SfMLearner[6]. We replaced this to the softplus function $d' = \mathrm{softplus}(d) = \log(1 + e^d)$. This ensures that the predicted disparity is always non-negative and eliminates the need for proper tuning of α and β for the target dataset.

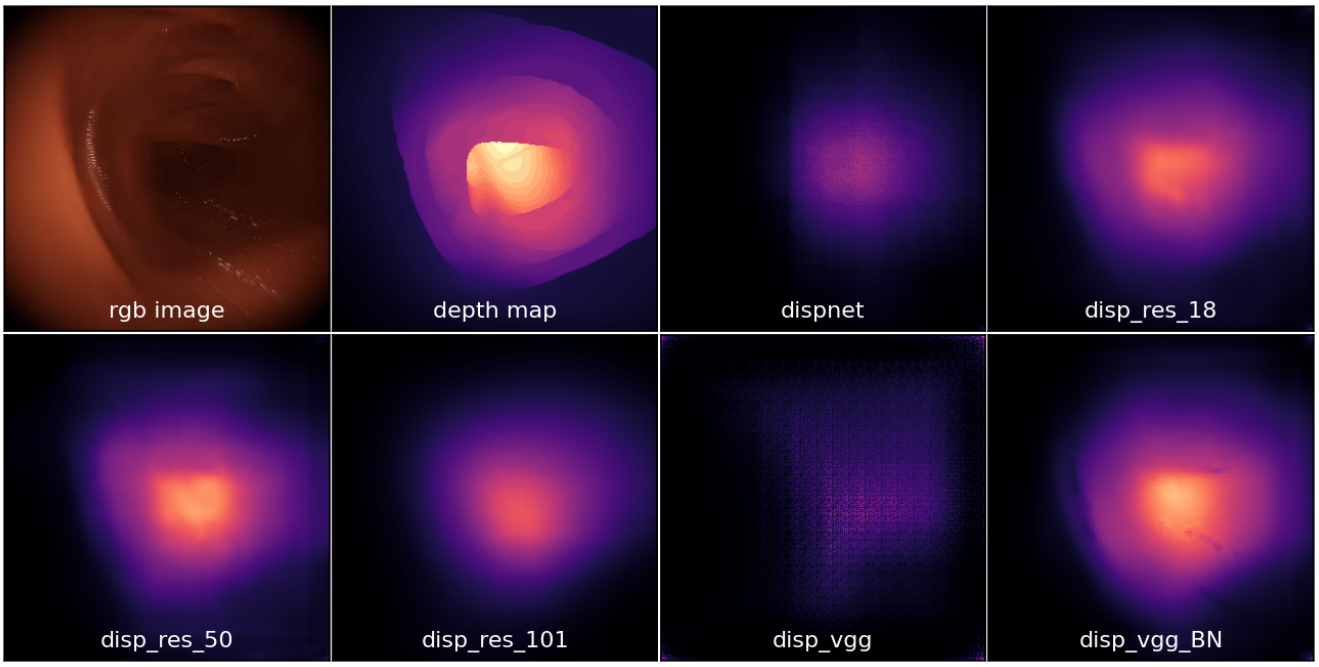

FIGURE 3 The trained network was applied to the RGB images in the validation dataset to predict the depth map. From top left to bottom right, the input RGB image, the truth depth map, and the predicted depth maps of the various trained networks. All depth map color maps are the same: black-magenta pixels are near and yellow-white pixels are far apart.

3.2 | Supervised Training of DispNet

For obtaining the scale-preserving monocular depth estimator, we train DispNet by optimizing the loss function comparing the predicted depth map and the ground truth depth map [14]. We used the synthetic dataset of synchronized sequential pairs of RGB images and depth maps for training DispNet (Sec. 2).

As for the loss for the supervised depth estimation, we used the simple L_1 loss:

$$E_{\text{supervised}} = \frac{1}{N} \sum_i |\tilde{D}_i - \hat{D}_i|, \tag{1}$$

where $\tilde{D}_i$ and $\hat{D}_i$ are the predicted and true depth values at all corresponding pixels, and N is the total number of the pixels.

We trained DispNet from scratch with the batch size of 4 and the data augmentation of random horizontal flip and random 90-degree rotations. RGB values were normalized so that RGB channels have a mean of $[0.5, 0.5, 0.5]$ and a standard deviation of $[0.5, 0.5, 0.5]$. The optimizer was Adam with the learning rate: 10^{-7}, momentum: 0.9, beta: 0.999, no weight decay. We used a cloud-based computing system called ABCI [17].

3.3 | Comparison of Encoders

Among the image encoders tested in [14], we used the following DispNet variants with five encoders: dispnet, disp_res_18, disp_res_50, disp_res_101, disp_vgg, disp_vgg_BN, which stand for the DispNet with the original DispNet [2,6], DispNet with ResNet-18/50/101, VGG without/with batch normalization, respectively.

The results of optimization are summarized in Tab. 1. The validation loss of the dispnet and disp_vgg were minimized in an early stage of the optimization process, which may mean that they are suffered from the over fitting. In terms of the MAE and the correlation coefficient, disp_vgg_BN is the best, but disp_res_50 and disp_res_18 follow with very minor differences.

The predicted depth maps are shown in Fig. 3. We can understand that the predictions of disp_res_18, disp_res_50 and disp_vgg_BN are visually close to the true depth map, which supports the previous insights of error metrics. When we carefully

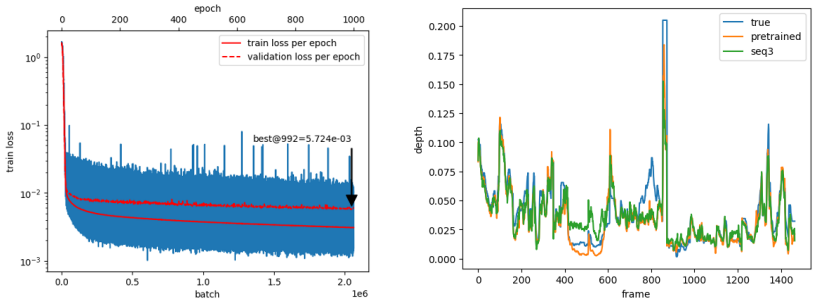

FIGURE 4 Left: plot of training and validation losses of DispNet with ResNet-50 encoder (disp_res_50). Right: The values of the true and predicted depth map at the image center (160, 160) through the sequence.

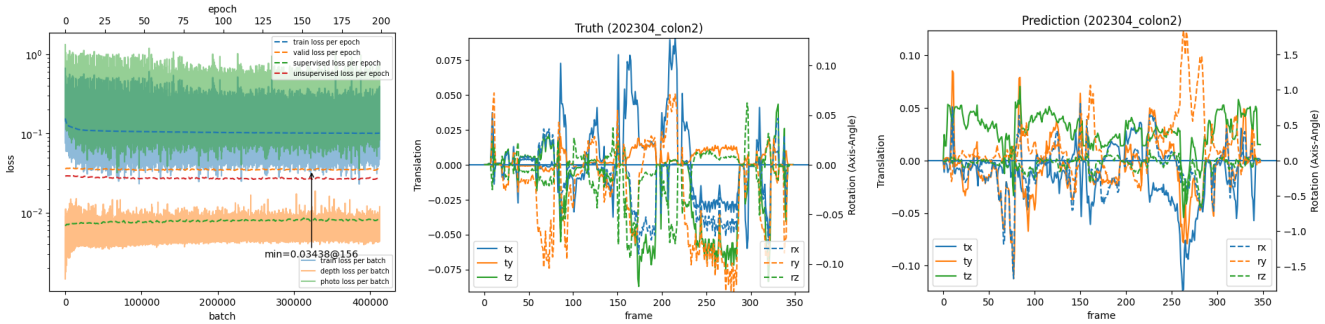

FIGURE 5 Training results of the PoseNet. Left: plot of training and validation losses during training. Middle and Right: the relative motion between successive frames.

inspect the predicted depth map, there are some noisy dots in disp_vgg_BN cases that don't exist in the ground truth depth map. Although it does not appear in Tab. 1, from a viewpoint of stability, we consider that ResNet based encoders (disp_res_18/_50) are more stable than VGG based encoders.

Figure 4 shows the supervised training procedure and the result of the DispNet with ResNet-50 encoder (disp_res_50). Comparison of the true and predicted depth values shows that the prediction of absolute depth values was successful.

4 | POSE ESTIMATION

4.1 | PoseNet

TABLE 1 Results for the VR-Caps validation dataset. The columns represent, respectively, the network model, the number of model parameters to train, the minimum training mean absolute error (MAE) and the number of epochs in which it occurred, the validation MAE (lower is better) and correlation coefficient (CC: higher is better), and training time. The best model is shown in bold and the second best model in italics. The learning epochs for DispNet was 500; the other models were 1000. The units for MAE are meters.

network	#params	least MAE	MAE	CC	time
dispnet	31596900	1.208e-2@111	0.01172	0.328	10:14
disp_res_18	14330244	5.955e-3@907	0.00568	0.872	20:29
disp_res_50	32523140	5.724e-3@992	*0.00540*	*0.873*	35:16
disp_res_101	62907268	7.584e-3@919	0.00728	0.782	57:58
disp_vgg	143507564	1.158e-2@30	0.01120	0.297	39:27
disp_vgg_BN	143516012	5.665e-3@919	**0.00539**	**0.883**	42:21

FIGURE 6 3D shape reconstruction results. The reconstructed shape from 50 frames of the true depth map (Left) and the predicted depth map (Middle), and the true poses and the estimated poses are compared (Right). The rotation parameters are in the axis-angle representation. Back faces are culled off for visualizing meshes.

PoseNet is a neural network that predicts a 6-DoF 3D relative rigid motion parameters (SE(3)) from a sequence of input images frames[18], and there are many varieties of implementations. For the target image, we take its preceding and succeeding images as the reference images to form a sequence. We used two types of PoseNet with different sequence length: one uses the target image and all the reference images as the input, and the other uses the target image and only one of the reference images, which are respectively referred by 'PoseExpNet' and 'PoseResNet' as appeared in the source codes of[6,14] and[8,12]. The former is implemented as a pure CNN with 1.6M parameters, and the latter uses ResNet18 as the feature detector followed by a CNN with 13M parameters in total.

There are also varieties of 6-DoF SE(3) pose parameterization used for SfMLearner descendants: a translation vector and a quaternion for rotation[6], a translation vector and Euler angles[12], a translation vector and a rotation vector (axis-angle representation or Rodrigues' formula)[10,7,8].

In the followings, we show the result of PoseExpNet that takes the ± 1-th images from the target image as the reference images, where the sequence length is 3, with the pose representation of the exponential map[19].

4.2 | Unsupervised Training of PoseNet

With the supervised DispNet, we train PoseNet in an unsupervised manner using its prediction. We used the loss function that is a weighted sum of the supervised and unsupervised losses:

$$E_{\text{total}} = w_{\text{sup}} E_{\text{sup}} + w_{\text{unsup}} E_{\text{unsup}},$$

where E_{sup} and E_{unsup} signify the supervised depth loss (Eq. (1)) and the unsupervised loss respectively. The unsupervised loss is the L_1 (the mean absolute error) of RGB values between the reprojected reference image and the target image, which is identical to the photometric consistency loss of[6,14]. The weights w_{sup} and w_{unsup} are for balancing these losses, and we set $w_{\text{sup}} = 1$ and $w_{\text{unsup}} = 1$.

We used DispNet with ResNet-50 encoders whose parameters were pre-trained in Sec. 3.2. We set the learning rates for DispNet and PoseNet to 10^{-7} and 10^{-4} respectively, and the optimizer settings are the same as Sec. 3.2. The optimization process was iterated for 200 epochs.

Figure 5 shows that the pose prediction is still imperfect and much more optimization is needed. The possible causes are the dataset, where the endoscope images are generally uniform and featureless, and our choice of PoseNet, pose parameterization and loss function.

5 | 3D SHAPE RECONSTRUCTION

With the trained PoseNet, we can predict the relative pose between the image frames, but the pose prediction is not reliable. In the original SfMLearner[6], the authors reported their pose estimation by the PoseNet outperformed the ORB-SLAM[9], but in many its variants, the output of PoseNet is refined by the ICP algorithm[12] or by the other SLAM methods like ORB-SLAM[9].

In our case, PoseNet did not predict the pose with sufficient quality, and we use the ICP algorithm[20] for more reliable shape reconstruction. Instead applying the ICP algorithm to successive frames, we use the global signed distance field (SDF) voxels as the global (canonical) shape, and align each frame to it[21,22]. We developed a SLAM algorithm 1 for reconstructing the global shape and the camera poses implemented by the Open3D Python packages[23]. The function 'convertToPointSet' converts a

Algorithm 1 Pseudocode for our SLAM algorithm

Require: $\{D_n\}$ $(0 \le n \le N-1)$, K ▷ A sequence of depth maps and the intrinsic matrix
1: $M_0 \leftarrow I_{4\times4}$ ▷ the pose of the first frame is fixed to the identity matrix
2: $V \leftarrow \emptyset$ ▷ set the global SDF voxels
3: **for** $n \leftarrow 0, N-1$ **do**
4: $P_n \leftarrow \text{convertToPointSet}(D_n, K)$ ▷ convert a depth map to a point set
5: **if** $n > 0$ **then**
6: $Q \leftarrow \text{extractPointSet}(V)$ ▷ extract a point set from the global SDF voxels
7: $M_n \leftarrow \text{ICP}(P_n, Q, M_{n-1})$ ▷ optimize the pose M_n from M_{n-1} such that $Q \approx M_n(P_n)$
8: **end if**
9: $V \leftarrow V \cup M_n(P_n)$ ▷ merge the transformed point set to the global SDF voxels
10: **if** $n > 0$ and $n \mod N_I = 0$ **then** ▷ global alignment at every $N_I = 50$ iterations
11: $Q \leftarrow \text{extractPointSet}(V)$ ▷ extract a point set from the global SDF voxels
12: **for** $m \leftarrow 1, n$ **do**
13: $M_m \leftarrow \text{ICP}(P_m, Q)$ ▷ update poses
14: **end for**
15: $V \leftarrow \emptyset$ ▷ reset the global SDF voxels
16: **for** $m \leftarrow 0, n$ **do**
17: $V \leftarrow V \cup M_m(P_m)$ ▷ update global SDF voxels
18: **end for**
19: **end if**
20: **end for**

depth map to a point set, where a depth value D at (u, v) of a depth map is converted to a 3D point by $\boldsymbol{p} = D \cdot K^{-1} \cdot (u, v, 1)^T$. The transformation $M(P)$ represents a 3D rigid motion whose rotation and translation are respectively R and $\boldsymbol{t}$, where a point $\boldsymbol{p} \in P$ is transformed to a point $\boldsymbol{p}'$ by $\boldsymbol{p}' = R\boldsymbol{p} + \boldsymbol{t}$. The function 'extractPointSet' and merging of point sets at lines 9 and 17 are implemented by calling Open3D functions.

Figure 6 shows the reconstruction result of this algorithm from the depth sequence of 50 frames predicted from the synthetic images. The predicted depth value is used only in the central region where pixels are located within the circle circumscribing the image boundary. The plots of poses estimated from the true depth is almost identical to the ground truth pose, and the 3D shape reconstructed from the predicted depth is similar to that from the true depth, which shows that the proposed algorithm worked correctly. For pose estimation from the predicted depth map, the rotation estimation is not accurate because the object shape is nearly cylindrical and no color information was used.

6 | APPLICATION TO REAL DATA

6.1 | Depth Prediction

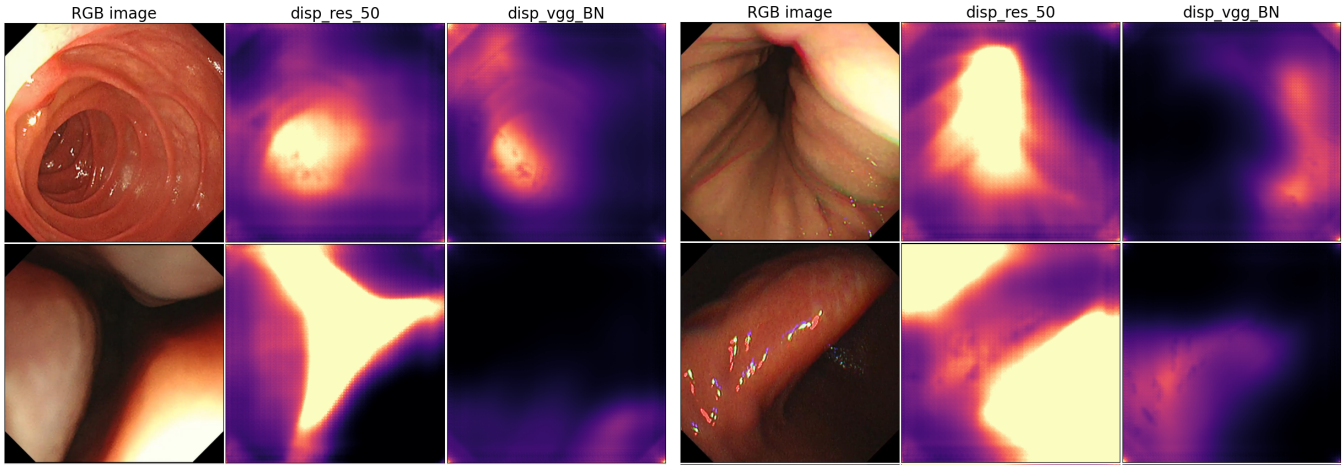

FIGURE 7 The example cases where evident difference can be observed in comparing the prediction of the DispNet with two different encoders: disp_res_50 and disp_vgg_BN. The depth colormap is common to all depth maps: black-magenta pixels are near and yellow-white pixels are far.

We applied the trained supervised DispNet to a real dataset that contains 13 RGB image sequences of digestive organs. There is no ground truth depth map that can be used for evaluation in this dataset. We estimated the camera parameters including the lens distortion from small portion of the sequence where the COLMAP[24] algorithm could work on, and rectify the images to adapt our DispNet. We chose two encoders: disp_res_50 and disp_vgg_BN for comparison, and we found that there are frames with significant difference between these two encoders: selected cases are shown in Fig. 7. In[14], the authors say that disp_vgg_BN is the best, and also in terms of the error metrics in Tab. 1, the performances of these encoders are similar, but considering the adaptation capability to the dataset that was not used for training, disp_res_50 works more stably than disp_vgg_BN. Due to lack of real endocsope images with ground truth depth map, we could not apply transfer learning for data adaptation.

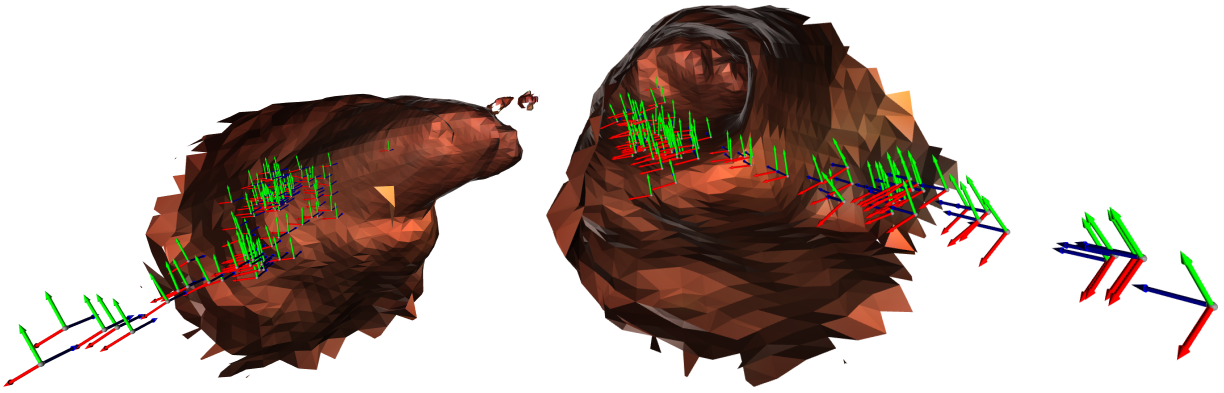

FIGURE 8 The reconstructed 3D shape and the camera poses rendered from two viewpoints. Back faces are culled off for visualizing meshes.

6.2 | 3D Shape Reconstruction

We applied our 3D reconstruction algorithm (Algorithm 1) on the predicted depth maps. Figure 8 shows the result of reconstruction from 110 images. Due to lack of the ground truth depth map, we cannot evaluate this result like as in Sec. 6.1, but this figure shows that a reasonable cylindrical 3D shape was reconstructed by the proposed algorithm.

7 | CONCLUSIONS

In this paper, we present supervised learning of depth prediction networks from monocular RGB images with absolute depth preservation. Synchronous sequences of RGB images and depth maps were generated using an endoscope simulator. The performance of various DispNet models was compared, and as far as we tested, the ResNet-based encoder performed the best. By training DispNet with the ground truth depth maps, the absolute depth maps were predicted from RGB images. According to interviews with clinicians, reconstructed shape accuracy should be at least 5 mm or less for practical use, and our results largely satisfied this requirement. We also developed a SLAM algorithm based on the ICP algorithm and SDF to align and integrate the predicted absolute depth maps to form a scale-preserving 3D shape model. These methods were applied to actual endoscopic image sequences.

We hope to improve our study to perform better in a variety of endoscopic scenes. Since it is difficult to obtain datasets of real endoscope RGB images with synchronized ground truth depth maps, we trained with synthetic datasets, but we need datasets with much varieties. For PoseNet, the performance for endoscopic images needs to be verified, since results of sufficient quality were not obtained. The proposed SLAM algorithm worked properly, but to improve its performance, it is necessary to use color and texture information in addition. Also, we would like to extend the SLAM algorithm to handle more realistic cases including deformations and turbulence.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP20H00611, JP18H04119, JP21H01457. This paper is based on results obtained from a project, JPNP20006, subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

1. Wang Y, Long Y, Fan SH, Dou Q. Neural Rendering for Stereo 3D Reconstruction of Deformable Tissues in Robotic Surgery. In: 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). 2022:431–441
2. Eigen D, Puhrsch C, Fergus R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K., eds. *Advances in Neural Information Processing Systems*. 27. Advances in Neural Information Processing Systems. Curran Associates, Inc. 2014.
3. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012:3354–3361
4. Silberman N, Hoiem D, Kohli P, Fergus R. Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C., eds. *Computer Vision – ECCV 2012*Computer Vision – ECCV 2012. Springer Berlin Heidelberg 2012; Berlin, Heidelberg:746–760.
5. Tateno K, Tombari F, Laina I, Navab N. CNN-SLAM: Real-Time Dense Monocular SLAM with Learned Depth Prediction. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society 2017; Los Alamitos, CA, USA:6565–6574
6. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised Learning of Depth and Ego-Motion from Video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:6612–6619
7. Wang C, Miguel Buenaposada J, Zhu R, Lucey S. Learning Depth From Monocular Videos Using Direct Methods. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.
8. Godard C, Mac Aodha O, Firman M, Brostow GJ. Digging into Self-Supervised Monocular Depth Prediction. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019
9. Bian JW, Zhan H, Wang N, et al. Unsupervised Scale-consistent Depth Learning from Video. *International journal of computer vision.* 2021. doi: 10.1007/s11263-021-01484-6
10. Bian JW, Zhan H, Wang N, Chin TJ, Shen C, Reid I. Auto-Rectify Network for Unsupervised Indoor Depth Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2021. doi: 10.1109/TPAMI.2021.3136220
11. Sun L, Bian JW, Zhan H, Yin W, Reid I, Shen C. SC-DepthV3: Robust Self-supervised Monocular Depth Estimation for Dynamic Scenes. https://doi.org/10.48550/arXiv.2211.03660; 2022.
12. Ozyoruk KB, Gokceler GI, Bobrow TL, et al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis.* 2021;71:102058. doi: https://doi.org/10.1016/j.media.2021.102058
13. İncetana K, Celikb IO, Obeida A, et al. VR-Caps: A Virtual Environment for Capsule Endoscopy. *Medical Image Analysis.* 2021;70:101990.
14. Fang Z, Chen X, Chen Y, Van Gool L. Towards Good Practice for CNN-Based Monocular Depth Estimation. In: Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020:1091–1100.
15. Haas JK. A history of the Unity game engine. 2014.
16. VR-Caps: A Virtual Environment for Capsule Endoscopy. https://github.com/CapsuleEndoscope/VirtualCapsuleEndoscopy.git; 2020.
17. AI Bridging Cloud Infrastructure (ABCI). https://docs.abci.ai/en/; .
18. Kendall A, Grimes M, Cipolla R. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In: IEEE International Conference on Computer Vision (ICCV). 2015:2938–2946.
19. Murray RM, Li Z, Sastry SS. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.
20. Besl PJ, McKay ND. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1992;14(2):239-256.

21. Masuda T. Registration and Integration of Multiple Range Images by Matching Signed Distance Fields for Object Shape Modeling. *Computer Vision and Image Understanding*. 2002;87(1-3):51-65.
22. Newcombe RA, Izadi S, Hilliges O, et al. KinectFusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality. 2011:127-136
23. Zhou QY, Park J, Koltun V. Open3D: A Modern Library for 3D Data Processing. *arXiv:1801.09847*. 2018.
24. Schönberger JL, Frahm JM. Structure-from-Motion Revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:4104–4113