Dense Depth Estimation from Stereo Endoscopy Videos Using Unsupervised Optical Flow Methods

Zixin Yang[1(✉)], Richard Simon[2], Yangming Li[3], and Cristian A. Linte[1,2]

[1] Center for Imaging Science, Rochester Institute of Technology, Rochester, NY 14623, USA
yy8898@rit.edu
[2] Biomedical Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA
[3] Electrical Computer and Telecommunications Engineering Technology, Rochester Institute of Technology, Rochester, NY 14623, USA

Abstract. In the context of Minimally Invasive Surgery, estimating depth from stereo endoscopy plays a crucial role in three-dimensional (3D) reconstruction, surgical navigation, and augmentation reality (AR) visualization. However, the challenges associated with this task are threefold: 1) feature-less surface representations, often polluted by artifacts, pose difficulty in identifying correspondence; 2) ground truth depth is difficult to estimate; and 3) an endoscopy image acquisition accompanied by accurately calibrated camera parameters is rare, as the camera is often adjusted during an intervention. To address these difficulties, we propose an unsupervised depth estimation framework (END-flow) based on an unsupervised optical flow network trained on un-rectified binocular videos without calibrated camera parameters. The proposed END-flow architecture is compared with traditional stereo matching, self-supervised depth estimation, unsupervised optical flow, and supervised methods implemented on the Stereo Correspondence and Reconstruction of Endoscopic Data (SCARED) Challenge dataset. Experimental results show that our method outperforms several state-of-the-art techniques and achieves a close performance to that of supervised methods.

Keywords: Stereo endoscopy · Depth estimation · Self supervised learning · Stereo matching · Optical flow

1 Introduction

In the context of Minimally Invasive Surgery (MIS), dense depth perception from endoscopy is a prerequisite for surgical robotics Augmented Reality (AR) [12] and computer vision-based navigation systems [23], as such applications require registration of pre-operative data, such as CT/MRI to intra-operative video data. Dense depth perception is also a fundamental component of simultaneous

B. W. Papież et al. (Eds.): MIUA 2021, LNCS 12722, pp. 337–349, 2021.
https://doi.org/10.1007/978-3-030-80432-9_26

localization and mapping (SLAM) [31] and three-dimension (3D) reconstruction [16]. However, estimating depth from endoscopic images is very challenging due to wet and feature-less surfaces, the presence of imaging artifacts, the presence of surgical instruments, and varying lighting conditions.

Depth can be estimated from different types of endoscopic images [7], including structured light endoscopy [14], monocular endoscopy [24] and stereo endoscopy [38]. By analyzing the deformation between the known projected light pattern and received projected pattern, structure light endoscopy can sparsely and accurately reconstruct tissue with no limitations due to texture information. Thus, structured light endoscopy is often used to reconstruct the ground truth depth [1,30]. However, this technique requires specialized processing hardware and is sensitive to environment lighting, limiting its application *in vivo*.

Recovering depth from monocular endoscopic video can be achieved via SLAM [3,20], Shape from Shading [28], and Structure from Motion (SfM) [18,37], as well as machine learning [24] and its integration with SfM [16]. However, monocular depth estimation is the most challenging and least accurate technique: not only does it require the estimation of the camera pose, which is difficult to obtain due to the lack of photometric constancy cross frames, but scale ambiguity is a common, inherent problem in monocular dense depth estimation. To mitigate these limitations, most efforts have been shifted to estimating depth from stereo endoscopy [1,7].

Estimating depth in stereo endoscopy can be achieved via densely matching pixels from a pair of binocular images. Following intrinsic and extrinsic camera calibration, matched points can be triangulated to recover depth using both classical and deep learning methods.

Classical methods include dense optical flow [5,26] and stereo matching methods [6,11], with the latter being the most common method in estimating depth [29]. Several traditional stereo matching methods have been applied [2,31,40] in MIS. However, despite achieving accurate results in the feature-rich region, stereo matching methods often lead to holes and speckle in texture-less surfaces, occlusions, repetition patterns, non-Lambertian surfaces, and specularities, which are common in endoscopy images. As such, parameter tuning and post-processing are necessary.

With the rapid development of deep learning, methods based on the convolutional neural network (CNN) have surpassed traditional methods in several public benchmarks, such as SceneFlow [21] and KITTI platform [22]. However, using CNNs in a supervised fashion in endoscopic videos is challenging, as ground truth depth is difficult to obtain. Visentini-Scarzanella *et al.* [33] trained and validated CNNs on phantom bronchus data from CT data. Similarly, Mahmood *et al.*[19] trained CNNs on synthetic texture-free images generated from digital colon phantom and validated on real images using adversarial learning to transfer real images to synthetic images. Lastly, Wang *et al.* [35] trained and validated a stereo matching CNN on simulated binocular data. A disparity dataset from CT [4] with a limited number of frames was created from *ex vivo* small porcine full torso cadavers and was used to assess several publicly supervised stereo matching methods.

Recently, self-supervised methods [9,39,42] that utilize image reconstruction as supervision signals have achieved remarkable results in self-driving cars. Their common approach is to formulate a depth estimation problem as the minimization of a photometric reprojection loss at the training stage. Self/unsupervised methods include self-supervised stereo matching [34], self-supervised depth estimation [9,39,42], and unsupervised optical flow [15]. However, they have been rarely studied on stereo endoscopic images.

To our best knowledge, the only self-supervised stereo matching method implemented on a pair of stereo endoscopic images was reported by Ye *et al.* in [38]. Nevertheless, there have been several works reported that focus on estimating monocular endoscopic image depth. In [16], Liu *et al.* incorporated recomputed matched points and camera pose from the SfM to train a self-supervised monocular depth estimation network on sinus video. Similarly, Ozyoruk *et al.* [24] jointly estimate camera pose and depth on synthetically generated data.

In the training stage, a self-supervised depth estimation network requires calibrated stereo camera parameters, while a self-supervised stereo matching requires rectified images. Nevertheless, both limit their application in MIS. During an intervention, the surgeon adjusts focus to adapt to anatomical targets, therefore invalidating pre-calibrated parameters [27], while binocular images can be rectified through an uncalibrated stereo rectification approach, which uses matched points to estimate the fundamental matrix. However, on a pair of featureless frames, accurately matched points are limited, and rectification error is introduced in this process [17]. Unsupervised optical flow methods have the advantage that, during training, they do not require calibrated camera parameters or an un-calibrated stereo rectification process.

To estimate depth from stereo endoscopic videos, we present an unsupervised optical flow network (END-flow). Compared with other methods, it does not require any ground truth labels, calibrated camera parameters, or rectified images for training. This work represents the first effort to use an unsupervised optical flow network to estimate depth from a stereo endoscopic video to the best of our knowledge. In addition, we also introduce an auto-masking and a sparse flow loss function to improve further accuracy beyond that achieved via the techniques disseminated to date.

2 Methods

The goal of this work is to first learn the optical flow mapping without the need of ground truth depth or camera parameters and subsequently recover depth from the optical flow with camera parameters in inference. Previous works [9, 15,39,41] have established solid baseline in unsupervised learning. In this section, we first introduce the general image reconstruction objective loss functions for unsupervised optical flow learning, then introduce our proposed enhancements in training using END-flow.

2.1 Baseline Unsupervised Optical Flow Loss Functions

In the absence of ground truth, one alternative is to use image reconstruction as the supervisory signal. The common approach is to formulate a photometric loss between the original image and the warped image. For two images, target image I_t and source image I_s, I_s can be warped to I_t via predicted optical flow mapping transformation $F_{t,s}$ to create the synthesis view of $I_t^{'}$ using

$$I_t^{'}(\boldsymbol{p}) = I_s(\boldsymbol{p} + F_{t,s}), \tag{1}$$

where $\boldsymbol{p}$ is the pixel coordinates on the target image. Following [9,39,41], the photometric loss can be established using

$$L_p = \frac{\alpha}{2}(1 - \mathrm{SSIM}(I_t, I_t^{'})) + (1 - \alpha)\|I_t - I_t^{'}\|_1, \tag{2}$$

where $\alpha = 0.85$, $SSIM$ is the similarity structure index [36]. However, the photometric loss may not be valid in texture-less regions; instead, at these locations, an edge-aware smoothness [8] term is commonly coupled with L_p, taking the form shown below:

$$L_s = |\partial_x F_{t,s}|\, e^{-|\partial_x I_t|} + |\partial_y F_{t,s}|\, e^{-|\partial_y I_t|}. \tag{3}$$

Overall, the total loss function for training via unsupervised optical flow mapping is of the form:

$$L_{flow} = L_p + \beta L_s, \tag{4}$$

where β is commonly set to 0.1.

2.2 Proposed Method

Previous works in optical flow networks take sequential images as input; here, we extend the method to find stereo correspondences on a pair of binocular images. The pipeline associated with the training stage is shown in Fig. 1. We adopt PWC-net [32] as our backbone network to predict forward flow, left to right, and backward flow, right to the left. The basic unsupervised loss function for a pair of binocular images is therefore given by:

$$L_{flow} = \sum_{I_t \in (I_l, I_r))} (L_p + \beta L_s). \tag{5}$$

Auto-masking. To handle occlusions and feature-less regions where photometric consistency is not valid, we utilize the auto-masking method proposed in [9] to select a valid region for photometric loss calculation:

$$M_p = \left[\, \|I_t - I_t^{'}\|_1 < \|I_t - I_{t'}\|_1 \,\right]. \tag{6}$$

Here [ ] is the Iverson bracket, taking the value 1 if the statement inside the bracket is true and otherwise taking the value 0. $I_{t'}$ is the other image in the pair of images, and $I_t^{'}$ is the synthesis image.

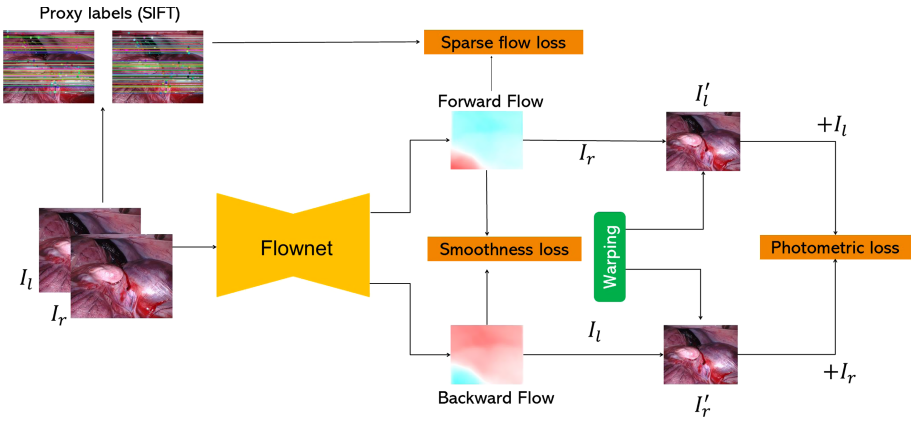

Fig. 1. Overview of proposed method. We adopt PWC-net [32] as our Flow-net. Proxy labels are generated from SIFT as a supervision signal. A sparse flow loss is calculated between proxy labels and predicted forward flow. Smoothness loss is calculated on forward flow and backward flow. The difference between the warped and input images forms the photometric loss.

Sparse Flow Loss. In addition to the basic unsupervised loss function, a sparse flow loss is included. We used an illumination-invariant feature descriptor, the Scale Invariant Feature Transform (SIFT), to find matched key-points within a pair of stereo images. Key-points are used to estimate the fundamental matrix, which is then used along with the RANSAC method to eliminate outliers. These matched points are further processed to generate a sparse flow map from the left image to the right image, to serve as a proxy label for supervision. The sparse flow F_{SIFT} loss is defined by:

$$L_{sf} = \frac{1}{|M_{SIFT}|} \sum_{p} M_{SIFT}(p) \| F_{SIFT}(p) - F_{l,r}(p_a) \|_1, \tag{7}$$

where $F_{l,r}$ is sparse flow map generated from SIFT, M_{SIFT} is the mask where sparse keypoints exist, and $|M_{SIFT}|$ stands for the number of matched points.

The overall loss function L is formulated as

$$L = \sum_{I_t \in (I_l, I_r))} (M_p L_p + \beta L_s) + \gamma L_{sf}, \tag{8}$$

where γ is the weight for sparse flow loss and is empirically set to 0.15.

To recover depth from optical flow in inference, following [10,41], we adopt the mid-point triangulation method using the stereo calibration parameters, which has a linear solution.

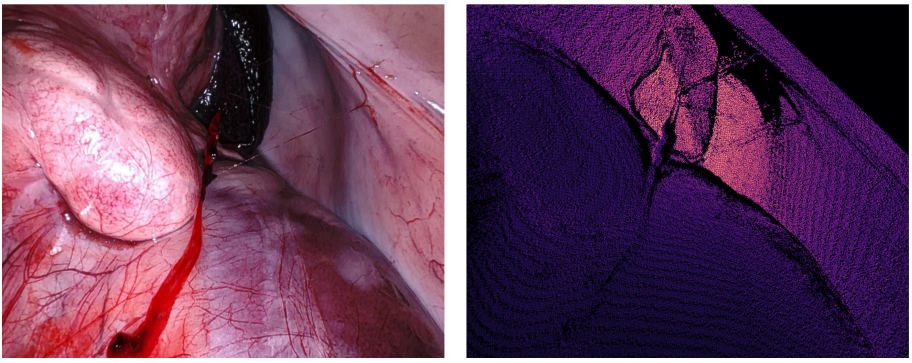

Fig. 2. An example of an endoscopic image, as well as its ground truth reconstructed depth from the SCARED dataset.

3 Dataset and Implementation

We conducted all experiments on the SCARED dataset[1] (the Stereo Correspondence and Reconstruction of Endoscopic Data). The dataset contains binocular images of abdominal anatomy from fresh porcine cadavers collected by a Da Vinci Xi endoscope, along with the associated camera parameters, camera poses, and ground truth depth maps generated using structure light. One sample is shown in Fig. 2. The data employed in this work consists of seven training datasets and two testing datasets.

All experiments are conducted on a GTX 2070 GPU, and all methods are implemented on Pytorch. We train the models [25] using the Adam optimizer [13] with a learning rate 10^{-4} and a batch size of 8. Images are enhanced with CLAHE (contrast limited adaptive histogram equalization) and resized to 256×320. Data augmentation only includes random flip, which also mimics the real scenario, especially in the event that the left - and right- images are flipped.

4 Results

4.1 Evaluation Metrics

We use the following metrics for evaluation: 1) the mean absolute distance (MAD (mm)), 2) the absolute relative error (AbsRel), and 3) the root mean squared error (RMSE (mm)), defined by the following equations:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{d}_i - d_i|, \tag{9}$$

$$AbsRel = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{d}_i - d_i|}{d_i}, \tag{10}$$

[1] https://endovissub2019-scared.grand-challenge.org/.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\hat{d}_i - d_i|^2}, \tag{11}$$

where n denotes the number of pixels, $\hat{d}_i$ and d_i represent ground truth depth and predicted depth of the pixel i, respectively.

4.2 Comparison with State-of-the-Art Depth Reconstruction Methods

Table 1. Comparison between several state-of-the-art depth stereo reconstruction methods and our proposed method (END-flow) in terms of Mean Absolute Distance (MAD), Absolute Relative Error (AbsRel) and Root-Mean-Squared Error (RMSE) in mean ± std. The statistical significance of the END-flow results against other methods is identified by $*(p < 0.005)$.

Method	MAD (mm)	AbsRel (%)	RMSE (mm)
* SGM [11]	9.37 ± 2.95	7.44 ± 3.47	8.37
* PASM [34]	16.87 ± 4.83	22.12 ± 4.78	20.72
* Monodepth2 [9]	7.03 ± 3.83	9.373 ± 4.514	9.02
* AR-flow [15]	6.65 ± 3.50	8.509 ± 3.965	9.40
END-flow	**5.40** ± 3.92	**7.17** ± 5.20	**7.55**

We first compare the results achieved using our proposed method to those obtained using several state-of-art methods, including the traditional stereo matching method SGM [11], unsupervised stereo matching method PASM [34], self-supervised depth estimation method Monodepth2 [9], unsupervised optical flow method AR-flow. Both SGM and PASM require rectified images as input, while Monodepth2 requires camera parameters. Our method, END-flow, does not require stereo rectification or camera parameters for training, which is advantageous in the endoscopy application. These results are summarized in Table 1.

It has been noted that the SCARED dataset was reported to have a calibration error [1]. After close examination, datasets 1–3 featuring minor calibration errors are used for comparison. We use the shortest video in each dataset for testing, the remaining for training and validation. In total, there are 7092 image pairs used for training, 787 image pairs used for validation, and 613 image pairs used for testing. The results in Table 1 suggest that our method achieves the best performances. The differences between the MAD errors from END-flow and other methods are statistical significance, characterized by $p < 0.005$.

Qualitative results are presented in Fig. 3. In comparison with other methods, SGM [11] fails to find correspondences in the ambiguous region, presenting black holes, while PASM [34] designed for rectified natural images with high dependence on epipolar constraints shows bad performance. Despite the fact that

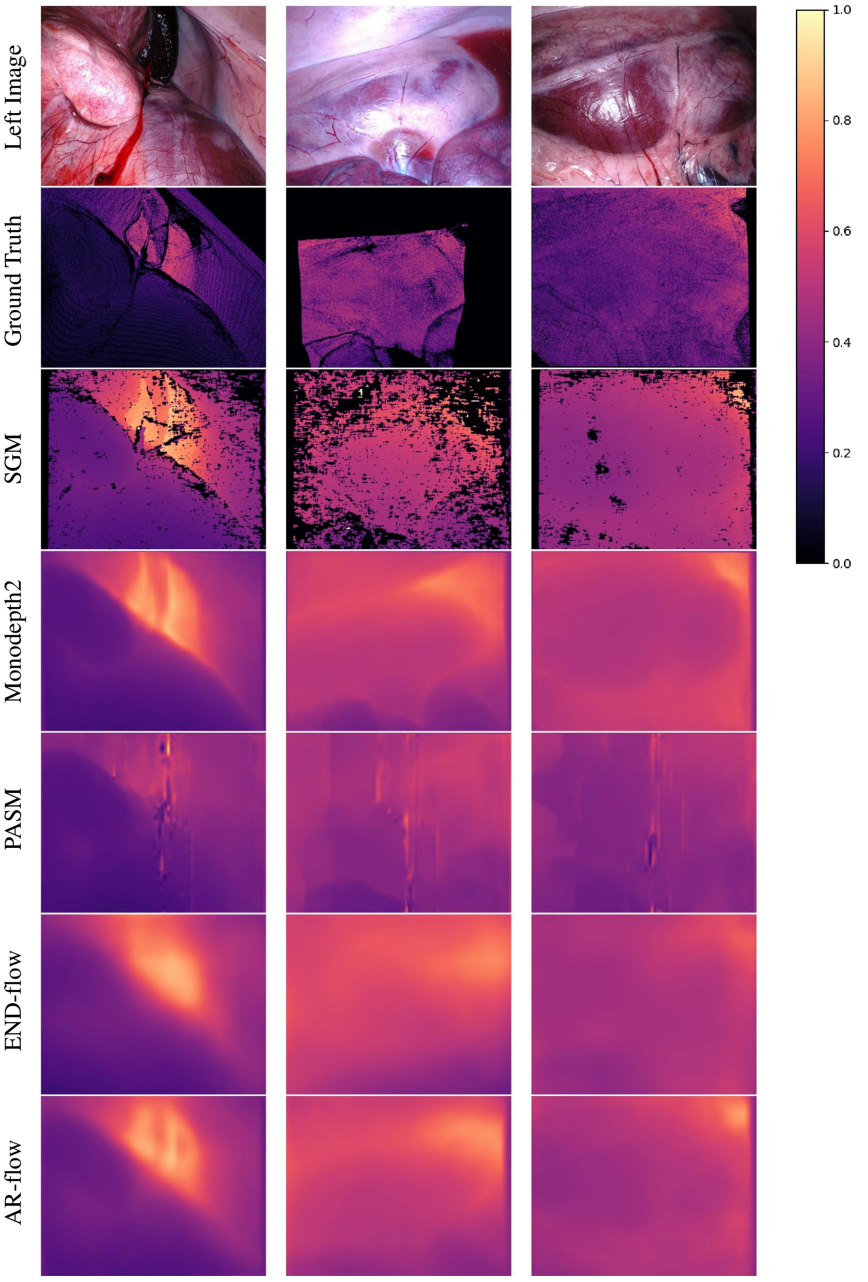

Fig. 3. Qualitative results achieved on three types of images (shown in 3 columns) as part of SCARED dataset using several techniques (illustrated in each row). Predicted depth maps are normalized by the maximum values of the ground truth depth map for enhanced visualization.

Monodepth2 predicts depth maps with sharp boundaries, our evaluation revealed that it tends to lose real scale. This may result from the fact that Monodepth2 takes one image to predict the depth, which is different from other methods that require two images. Moreover, over-enhanced edges on some images may result from texture changes, not necessarily depth changes. As such, the Monodepth2 technique may lead to over-enhanced edges when estimating depth from one image. Our method, END-flow, predicts depth with finer detail than AR-flow, while not lose real scale as Monodepth2.

With calibrated camera parameters, images can be rectified. When training or make predictions based on rectified images, optical flow can be constrained to the horizontal direction, the disparity, and this approach may help improve depth estimation. However, endoscopy image acquisition accompanied by accurately calibrated camera parameters is rare, as the camera is often adjusted during an intervention. To mitigate this inconvenience, our proposed method, END-flow, has the advantage of not requiring accurately calibrated camera parameters for training.

4.3 Ablation Study

To identify the contribution brought forth by each of the individual components integrated into our proposed pipeline, specifically auto-masking M_p and sparse flow loss L_{sf}, we conduct an ablation study to evaluate the performance of each pipeline component. This study is summarized in Table 2. Both M_p and L_{sf} alone, as well as their combination $M_p + L_{sf}$ yield statistically significant improvement in MAD compared to the baseline ($p < 0.005$).

Table 2. Ablation study showing the improvement in MAD (mm) (mean±std) in response to augmenting the baseline technique with the auto-masking M_p, sparse flow loss L_{sf} and their combination $M_p + L_{sf}$.

Method	MAD (mm)	AbsRel (%)	RMSE (mm)
Baseline	8.40 ± 4.08	11.02 ± 4.60	11.38
Baseline + M_p	7.46 ± 4.25	9.87 ± 4.95	10.20
Baseline + L_{sf}	5.59 ± 3.92	7.37 ± 5.17	7.76
Baseline + M_p + L_{sf}	5.40 ± 3.92	7.17 ± 5.20	7.55

4.4 Comparison with Top Methods in the SCARED Challenge

We further compare our method with winners' methods reported in the SCARED challenge [1], shown in Table 3. Winners were Trevor Zeffiro and Jean-Claude Rosenthal. We train our method on seven training sub-datasets and test on two testing sub-datasets. Note that these winners' methods utilized ground truth depth for training their networks, while our proposed architecture method achieves competitive results without using the ground truth depth labels.

Table 3. Comparison (in terms of mean MAD (mm)) between END-flow and best performing methods reported in the SCARED challenge.

	Trevor Zeffiro [1]	J.C. Rosenthal [1]	END-flow
testDataset1	3.60	3.44	4.77
testDatseta2	3.47	4.05	4.76

5 Conclusion

We have presented a dense depth estimation method based on an unsupervised optical flow network named END-flow. This method poses several advantages over previous techniques: 1) it can be trained on original videos without access to camera calibration parameters and stereo rectification or ground-truth labels; and 2) it integrates key-points matching to facilitates training. We deployed this method on several datasets available as part of the SCARED challenge; the results achieved using END-flow are comparable to those achieved using state-of-art methods, as well as the best-performing methods reported in the challenge. Specifically, we demonstrate that END-flow outperforms the state-of-the-art traditional and self/unsupervised methods and achieves comparatively performance against the best-performing supervised methods reported in the challenge. Future work will focus on estimating the confidence of unsupervised optical flow methods, which will benefit the down-stream analysis and integration of traditional depth estimation methods. Following additional work on the topic and further software improvement, we plan to release a link to a repository consisting of open-source code to the community.

References

1. Allan, M., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. Bernhardt, S., Abi-Nahed, J., Abugharbieh, R.: Robust dense endoscopic stereo reconstruction for minimally invasive surgery. In: Menze, B.H., Langs, G., Lu, L., Montillo, A., Tu, Z., Criminisi, A. (eds.) MCV 2012. LNCS, vol. 7766, pp. 254–262. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36620-8_25
3. Chen, L., Tang, W., John, N.W., Wan, T.R., Zhang, J.J.: Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. Comput. Methods Prog. Biomed **158**, 135–146 (2018)
4. Eddie"Edwards, P., Psychogyios, D., Speidel, S., Maier-Hein, L., Stoyanov, D.: Serv-ct: a disparity dataset from ct for validation of endoscopic 3d reconstruction. arXiv e-prints pp. arXiv-2012 (2020)
5. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45103-X_50
6. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6492, pp. 25–38. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19315-6_3

7. Geng, J., Xie, J.: Review of 3-d endoscopic surface imaging techniques. IEEE Sens. J. **14**(4), 945–960 (2013)
8. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270–279 (2017)
9. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3828–3838 (2019)
10. Hartley, R.I., Sturm, P.: Triangulation. Comput. Vision Image Underst. **68**(2), 146–157 (1997)
11. Hirschmuller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, pp. 807–814. IEEE (2005)
12. Kalia, M., Navab, N., Salcudean, T.: A real-time interactive augmented reality depth estimation technique for surgical robotics. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 8291–8297. IEEE (2019)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Lin, J., et al.: Endoscopic depth measurement and super-spectral-resolution imaging. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 39–47. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_5
15. Liu, L., et al.: Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6489–6498 (2020)
16. Liu, X., et al.: Reconstructing sinus anatomy from endoscopic video – towards a radiation-free approach for quantitative longitudinal assessment. In: Martel, A.L., Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 3–13. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_1
17. Luo, X., Jayarathne, U.L., McLeod, A.J., Pautler, S.E., Schlacta, C.M., Peters, T.M.: Uncalibrated stereo rectification and disparity range stabilization: a comparison of different feature detectors. In: Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 9786, p. 97861C. International Society for Optics and Photonics (2016)
18. Lurie, K.L., Angst, R., Zlatev, D.V., Liao, J.C., Bowden, A.K.E.: 3d reconstruction of cystoscopy videos for comprehensive bladder records. Biomed. Opt. Exp. **8**(4), 2106–2123 (2017)
19. Mahmood, F., Durr, N.J.: Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Med. Image Anal. **48**, 230–243 (2018)
20. Mahmoud, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., Montiel, J.M.M.: Live tracking and dense reconstruction for handheld monocular endoscopy. IEEE Trans. Medical Imag. **38**(1), 79–89 (2018)
21. Mayer, N., et al.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4040–4048 (2016)
22. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3061–3070 (2015)
23. Mirota, D.J., Ishii, M., Hager, G.D.: Vision-based navigation in image-guided interventions. Ann. Rev. Biomed. Eng. **13** (2011)

24. Ozyoruk, K.B., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med. Image Anal., 102058 (2021)
25. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
26. Phan, T.B., Trinh, D.H., Lamarque, D., Wolf, D., Daul, C.: Dense optical flow for the reconstruction of weakly textured and structured surfaces: Application to endoscopy. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 310–314. IEEE (2019)
27. Pratt, P., Bergeles, C., Darzi, A., Yang, G.Z.: Practical intraoperative stereo camera calibration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 667–675. Springer (2014)
28. Ren, Z., He, T., Peng, L., Liu, S., Zhu, S., Zeng, B.: Shape recovery of endoscopic videos by shape from shading using mesh regularization. In: Zhao, Y., Kong, X., Taubman, D. (eds.) ICIG 2017. LNCS, vol. 10668, pp. 204–213. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71598-8_19
29. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision **47**(1), 7–42 (2002)
30. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, Proceedings, vol. 1, pp. I-I. IEEE (2003)
31. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Mis-slam: real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. IEEE Rob. Autom. Lett. **3**(4), 4068–4075 (2018)
32. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8934–8943 (2018)
33. Visentini-Scarzanella, M., Sugiura, T., Kaneko, T., Koto, S.: Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. Int. J. Comput. Assist. Radiol. Surg. **12**(7), 1089–1099 (2017)
34. Wang, L., et a.: Parallax attention for unsupervised stereo correspondence learning. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
35. Wang, X.Z., Nie, Y., Lu, S.P., Zhang, J.: Deep convolutional network for stereo depth mapping in binocular endoscopy. IEEE Access **8**, 73241–73249 (2020)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process **13**(4), 600–612 (2004)
37. Widya, A.R., Monno, Y., Okutomi, M., Suzuki, S., Gotoda, T., Miki, K.: Whole stomach 3D reconstruction and frame localization from monocular endoscope video. IEEE J. Transl. Eng. Health Med. **7**, 1–10 (2019)
38. Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.Z.: Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. In: Hamlyn Symposium on Medical Robotics (2017)
39. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1983–1992 (2018)
40. Zampokas, G., Tsiolis, K., Peleka, G., Mariolis, I., Malasiotis, S., Tzovaras, D.: Real-time 3D reconstruction in minimally invasive surgery with quasi-dense matching. In: 2018 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6. IEEE (2018)

41. Zhao, W., Liu, S., Shu, Y., Liu, Y.J.: Towards better generalization: joint depth-pose learning without posenet. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9151–9161 (2020)
42. 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)