Depth Estimation for Colonoscopy Images with Self-supervised Learning from Videos

Kai Cheng[1], Yiting Ma[1], Bin Sun[3], Yang Li[3], and Xuejin Chen[1,2](✉)

[1] National Engineering Laboratory for Brain-inspired Intelligence Technology and Application, University of Science and Technology of China, Hefei, China
xjchen99@ustc.edu.cn
[2] Institute of Artificial Intelligence, Hefei, China
[3] The First Affiliated Hospital of Anhui Medical University, Hefei, China

Abstract. Depth estimation in colonoscopy images provides geometric clues for downstream medical analysis tasks, such as polyp detection, 3D reconstruction, and diagnosis. Recently, deep learning technology has made significant progress in monocular depth estimation for natural scenes. However, without sufficient ground truth of dense depth maps for colonoscopy images, it is significantly challenging to train deep neural networks for colonoscopy depth estimation. In this paper, we propose a novel approach that makes full use of both synthetic data and real colonoscopy videos. We use synthetic data with ground truth depth maps to train a depth estimation network with a generative adversarial network model. Despite the lack of ground truth depth, real colonoscopy videos are used to train the network in a self-supervision manner by exploiting temporal consistency between neighboring frames. Furthermore, we design a masked gradient warping loss in order to ensure temporal consistency with more reliable correspondences. We conducted both quantitative and qualitative analysis on an existing synthetic dataset and a set of real colonoscopy videos, demonstrating the superiority of our method on more accurate and consistent depth estimation for colonoscopy images.

Keywords: Colonoscopy · Depth estimation · Self-supervised learning · Videos · Temporal consistency

1 Introduction

Colorectal cancer is recently reported as the third most prevalent malignancy and the fourth most common cause of cancer-associated death worldwide [1,13,15]. Colonoscopy is an effective technique for the prevention and treatment of

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87231-1_12) contains supplementary material, which is available to authorized users.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12906, pp. 119–128, 2021.
https://doi.org/10.1007/978-3-030-87231-1_12

colon cancer. Many approaches have been proposed for colorectal polyp detection and diagnosis in colonoscopy images and videos [5,9,20,21]. While geometric features, e.g. location, size, and shape of polyps, are critical for colorectal polyp diagnosis, depth estimation from colonoscopy images could help a lot in deriving 3D geometric information of the intestinal environment.

Many efforts have been put into depth estimation and 3D reconstruction of intestinal environments from colonoscopy videos. Low-level geometric clues are utilized in earlier model-based approaches. Hong *et al.* [3] estimate depths from colon fold contours. Zhao *et al.* [22] combine structure-from-motion (SfM) and shape-from-shading (SfS) techniques for surface reconstruction. These approaches suffer from reflection and low texture of colonoscopy images, resulting in inconsistency and great sparseness of the estimated depth maps. In order to enhance surface textures for more robust SfM, Widya *et al.* [17–19] use chromoendoscopy with the surface dyed by indigo carmine. However, chromoendoscopy is not very common, which leads to limitations in application. Besides, these approaches require dense feature extraction and global optimization, which is computationally expensive.

Deep-learning-based approaches have achieved remarkable performance in general depth estimation recently. Compared with natural scenes where ground-truth depth can be obtained using depth cameras or LiDARs, acquiring the ground-truth depth for colonoscopy videos is arduous. Ma *et al.* [7] use an SfM approach [12] to generate sparse colonoscopy depth maps as ground-truth to train a depth estimation network. However, due to the inherited limitation of SfM in low-quality reconstruction for textureless and non-Lambertian surfaces, it is challenging to obtain accurate dense depth maps for supervised learning. Assuming temporal consistency between frames in videos, unsupervised depth estimation has also been studied [2,6,23]. Liu *et al.* [6] propose a self-supervised depth estimation method for monocular endoscopic images using depth consistency check between adjacent frames with camera poses estimated by SfM. Freedman *et al.* [2] propose a calibration-free unsupervised method by predicting depth, camera pose, and intrinsics simultaneously. However, for colonoscopy videos with weak illumination in complex environments, these unsupervised approaches face significant challenges posed by frequent occlusions between colon folds and non-Lambertian surfaces.

Many works use synthetic data to produce precise ground truth depth for network training. Mahmood *et al.* [8] train a joint convolutional neural network-conditional random field framework on synthetic data and transfer real endoscopy images to synthetic style using a transformer network. Rau *et al.* [11] train an image translation network pix2pix [4] with synthetic image-and-depth pairs to directly translate a colonoscopy image into a depth map. In order to reduce the domain gap between synthetic data and real images, the GAN loss also involves the depth maps predicted from real colonoscopy images but L_1 loss is not computed since no ground truth is available for real images. By doing so, the generator is expected to learn to predict realistic-looking depth maps from real images. However, without accurate supervision on the details in the predicted depth map, it is non-trivial for the generator to precisely predict depth for unseen textures in real colonoscopy images that deviate from synthetic data.

In this paper, we not only utilize synthetic data with ground truth depth to help the network learn fine appearance features for depth estimation but also exploit the temporal consistency between neighboring frames to make full use of unlabeled real colonoscopy videos for self-supervision. Moreover, we design a masked gradient warping loss to filter out non-reliable correspondence caused by occlusions or reflections. A more powerful image translation model [16] is also employed in our framework to enhance the quality of depth estimation. We evaluate our method on the synthetic dataset [11] and our real colonoscopy videos. The results show that our method achieves more accurate and temporally consistent depth estimation for colonoscopy images.

2 Methodology

Given a single colonoscopy image $\mathbf{F}$, our goal is to train a deep neural network DepthNet G that directly generates a depth map $\mathbf{D}$ as $\mathbf{D} = G(\mathbf{F})$. In order to train the DepthNet G, we leverage both the synthetic data for full supervision and real colonoscopy videos for self-supervision via temporal consistency. The framework of our approach is shown in Fig. 1. First, we adopt a high-resolution image translation model to train DepthNet in an adversarial manner with synthetic data. Second, we introduce self-supervision during the network training by enforcing temporal consistency between the predicted depths of neighboring frames of real colonoscopy videos.

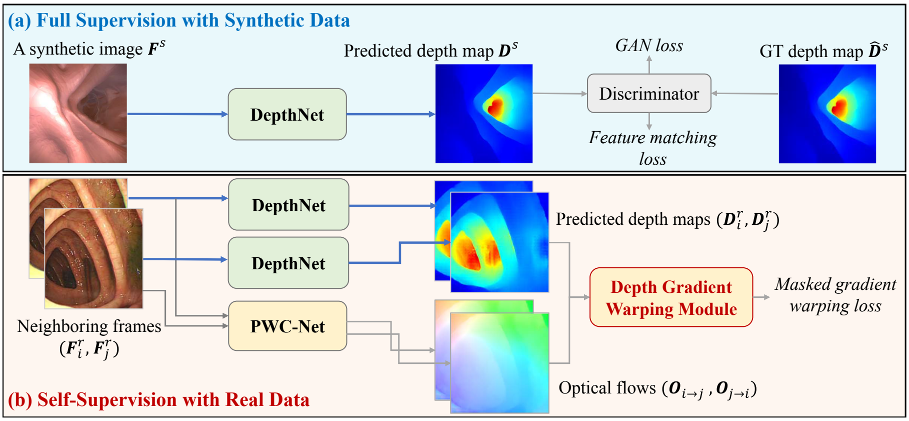
(a) Full Supervision with Synthetic Data
A synthetic image F^s
DepthNet
Predicted depth map D^s
GAN loss
Discriminator
Feature matching loss
GT depth map $\hat{D}^s$
DepthNet
DepthNet
PWC-Net
Predicted depth maps (D_i^r, D_j^r)
Depth Gradient Warping Module
Masked gradient warping loss
Neighboring frames (F_i^r, F_j^r)
Optical flows ($O_{i\to j}, O_{j\to i}$)
(b) Self-Supervision with Real Data

Fig. 1. Overview of our approach. (a) We first train DepthNet as a conditional GAN with synthetic image-and-depth pairs. (b) The DepthNet is then finetuned with self-supervision by checking the temporal consistency between neighboring frames.

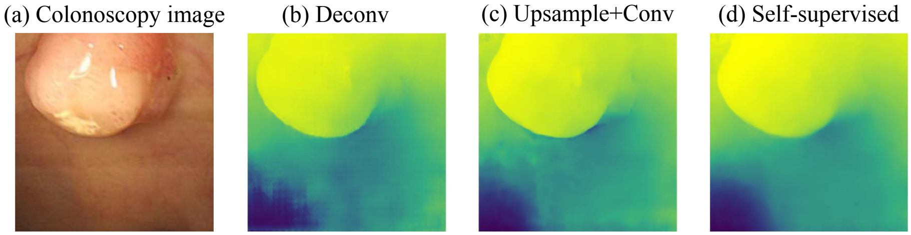

Fig. 2. Checkerboard artifacts. From a colonoscopy image (a), the original pix2pixHD model produces a depth map with checkerboard artifacts (b). Checkerboard artifact is alleviated by replacing the deconvolution layers in the generator with upsampling and convolution layers (c). Smoother depth is generated with our self-supervised model (d).

2.1 Training Baseline Model with Synthetic Data

We adopt the high-resolution image translation network pix2pixHD [16] as our baseline model to translate a colonoscopy image to a depth map. It consists of a coarse-to-fine generator and a multi-scale discriminator in order to produce high-resolution images. The network is trained in an adversarial manner with a GAN loss and feature matching loss [16] on the synthetic dataset [11] which contains paired synthetic colonoscopy images and the corresponding depth maps. However, the original pix2pixHD model produces results with checkerboard artifacts [10], as Fig. 2(b) shows. In order to alleviate this effect, we replace the deconvolution layers in the generator with upsampling and convolutional layers, similar to [6]. Figure 2(c) shows that the checkerboard effect is alleviated by replacing the deconvolutional layers with upsampling and convolutional layers. However, there are still many noises in the predicted results due to the specular reflections and textures, which appear frequently in real colonoscopy images.

2.2 Self-supervision with Colonoscopy Videos

Due to the domain gap between synthetic and real colonoscopy images, when applying the DepthNet trained on the synthetic data to predict depth directly from clinical colonoscopy images, the results tend to be spatially jumping and temporally inconsistent because of the specular reflection and complex textures in intestinal environments, as Fig. 2(c) shows. While obtaining ground-truth depth for real colonoscopy images is arduous, the temporal correlation between neighboring frames in colonoscopy videos provides natural constraints on the predicted depths. Therefore, we propose to enforce temporal consistency between the predicted depths of neighboring frames in network training.

For two neighboring frames in a real colonoscopy video $\mathbf{F}_i^r$ and $\mathbf{F}_j^r$, the DepthNet estimates two depth maps $\mathbf{D}_i^r$ and $\mathbf{D}_j^r$ respectively. In order to check the consistency between these two depth maps, a typical way is to warp one frame to the other according to the camera pose and intrinsic, which are not easy to obtain. In order to avoid camera calibration, we propose a calibration-free warping module that finds pixel correspondences from optical flows. A pre-trained

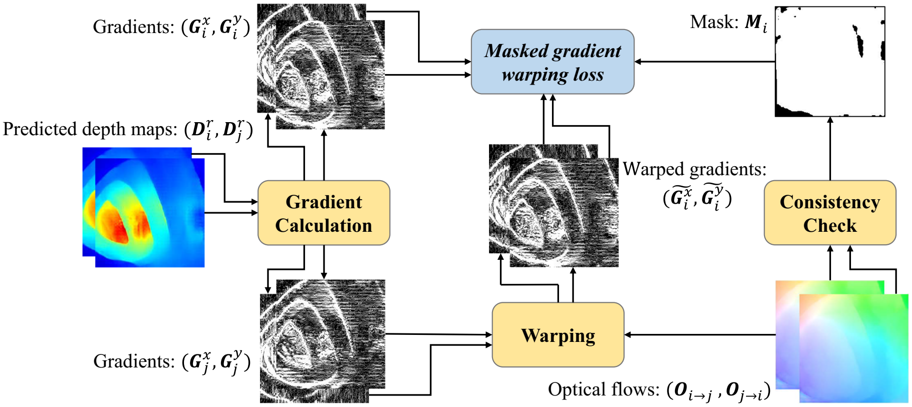
Gradients: (G_i^x, G_i^y)
Masked gradient warping loss
Mask: M_i
Predicted depth maps: (D_i^r, D_j^r)
Warped gradients: $(\widetilde{G_i^x}, \widetilde{G_i^y})$
Gradient Calculation
Consistency Check
Warping
Gradients: (G_j^x, G_j^y)
Optical flows: $(O_{i\to j}, O_{j\to i})$

Fig. 3. Depth gradient warping module to check the temporal structural consistency of the predicted depth maps of two neighboring frames.

network PWC-Net [14] is employed to infer optical flows. Due to self-occlusions and reflections in colons, brightness consistency is not guaranteed so that errors in optical flow estimation are inevitable. In order to filter out the optical flow noises, we estimate optical flows $\mathbf{O}_{i\to j}$ and $\mathbf{O}_{j\to i}$ in two directions. Then we check if a pixel $\mathbf{p}$ can be warped back to the same position from frame i to frame j by $\mathbf{O}_{i\to j}$ then from frame j to frame i by $\mathbf{O}_{j\to i}$. If not, the pixel $\mathbf{p}$ is filtered out when checking temporal consistency. Therefore, we compute a mask $\mathbf{M}_i$ for frame i as

$$\mathbf{M}_i(\mathbf{p}) = \begin{cases} 0, & |\mathbf{O}_{i\to j}(\mathbf{p}) + \mathbf{O}_{j\to i}(\mathbf{q})| > \varepsilon \\ 1, & otherwise \end{cases} \tag{1}$$

where $\mathbf{q}$ is the corresponding location in frame $\mathbf{F}_j^r$ of the pixel $\mathbf{p}$ in frame $\mathbf{F}_i^r$ according to the estimated optical flow $\mathbf{q} = \mathbf{p} + \mathbf{O}_{i\to j}(\mathbf{p})$. Note that we use bilinear interpolation of $\mathbf{O}_{j\to i}(\mathbf{q})$ for a subpixel $\mathbf{q}$. ε is a threshold for the forward-backward warping distance check. We set $\varepsilon = 1$ in our experiments.

However, the camera shifts at two neighboring frames. As a result, the absolute depth values of the correspondence pixels in two neighboring frames are not equal. Instead of comparing the depth values directly, we encourage the structural consistency between two depth maps by comparing the gradients of two depth maps through the depth gradient warping module. As Fig. 3 shows, we compute the gradients $(\mathbf{G}_i^x, \mathbf{G}_i^y)$, $(\mathbf{G}_j^x, \mathbf{G}_j^y)$ of the two predicted depth maps $\mathbf{D}_i^r$ and $\mathbf{D}_j^r$ in x and y direction. Then we check the consistency between the depth gradients of two neighboring frames with the mask $\mathbf{M}_i$ to calculate the masked gradient warping loss for self-supervision:

$$L_{MGW} = \frac{1}{|\mathbf{M}_i|} \sum_{\mathbf{p}\in\mathbf{F}_i^r} \mathbf{M}_i(\mathbf{p}) \Big(\Big|\mathbf{G}_i^x(\mathbf{p}) - \widetilde{\mathbf{G}}_i^x(\mathbf{p})\Big| + \Big|\mathbf{G}_i^y(\mathbf{p}) - \widetilde{\mathbf{G}}_i^y(\mathbf{p})\Big| \Big), \tag{2}$$

where $\widetilde{\mathbf{G}}_i^x, \widetilde{\mathbf{G}}_i^y$ are the gradient maps warped from $\mathbf{G}_j^x, \mathbf{G}_j^y$ according to the estimated optical flow $\mathbf{O}_{j \to i}$ by bilinear interpolation.

Our full objective combines both self-supervision with masked gradient warping loss L_{MGW} and supervision with GAN loss L_{GAN} and feature matching loss L_{FM} with α and γ balance the three loss terms:

$$L = \alpha L_{MGW} + \gamma L_{FM} + L_{GAN}. \tag{3}$$

3 Experiments

3.1 Dataset and Implementation Details

Both synthetic and real colonoscopy data are used for training and evaluation. We use the UCL synthetic dataset published by Rau *et al.* [11]. The dataset consists of 16,016 pairs of synthetic endoscopic images and the corresponding depth maps. Following their split strategy, the dataset is divided randomly into training, validation, and test set by 6:1:3. We also collect 57 clinical colonoscopy videos from different patients. In the training stage, we use neighboring frames from each video at different intervals. Trading off overlap and interval between frame pairs, we choose four intervals including 1, 4, 8, and 16 frames. The final dataset of real colonoscopy data contains 6,352 training pairs and 4,217 test pairs.

Both the synthetic images and real images are resized to 512 × 512. We train our network in two steps. In the first step, we train our model on the synthetic data only. In the second step, we finetune the model with self-supervision on real colonoscopy frames. The batch size of synthetic images and real images for the first step and second step is set 8 and 4 respectively. We employ Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate starts with $5e - 5$ and linearly decays. We update the generator every iteration while update the discriminator every 5 iterations. The framework is implemented in PyTorch 1.4 and trained on 4 Nvidia Titan XP GPUs. The first step training takes 70 epochs and we add the second step finetuning with real data at the last 10 epochs. The weight for the masked gradient warping loss $\alpha = 5$ initially and linearly increases by 2 in the second step. The weight of feature matching loss $\gamma = 2$.

3.2 Quantitative Evaluation

In order to quantitatively evaluate the performance of our method on depth estimation of colonoscopy images, we compare our method with previous approaches on the UCL synthetic dataset [11]. We adopt the same three metrics including the absolute L_1 distance, the relative error, and the root-mean-squared-error $RMSE$ between the ground truth and prediction. The results are reported in Table 1.

Table 1. Quantitative evaluation on the UCL synthetic dataset (* in cm, ** in %).

Method	Mean L_1-error*	Mean relative L_1-error**	Mean RMSE*
Pix2pix [11]	0.232 ± 0.046	8.2 ± 2.0	0.236
Extended pix2pix [11]	0.171 ± 0.034	6.4 ± 1.7	0.175
Our baseline	0.032 ± 0.011	1.5 ± 2.0	0.056
Ours	0.033 ± 0.012	1.6 ± 2.2	0.057

Our baseline model is only trained with synthetic data. It shows that a better conditional GAN model (pix2pixHD instead of pix2pix) brings great performance improvement. While only tested on synthetic data, our model that is fine-tuned with real colonoscopy videos does not make further improvement on synthetic data. This is reasonable because the self-supervision between neighboring frames leverages temporal consistency for more depth data from real colonoscopy video but it does not bring more information for the synthetic data.

Although the self-supervision with temporal consistency does not bring gain on the mean accuracy, it significantly improves the temporal consistency between the estimated depths on both the synthetic and real colonoscopy data. We quantify the temporal consistency by the masked gradient warping loss L_{MGW}, which reflects the structural consistency between the estimated depth maps of two neighboring frames. Table 2 demonstrates that our method reduces the masked gradient warping loss on both the synthetic data and real data.

Table 2. Masked gradient warping loss on synthetic and real colonoscopy datasets.

Method	Synthetic dataset (cm/pixel)	Real dataset (cm/pixel)
Baseline	0.024 ± 0.003	0.025 ± 0.020
Ours	$\mathbf{0.020 \pm 0.003}$	$\mathbf{0.012 \pm 0.011}$

3.3 Qualitative Evaluation on Real Data

Without ground-truth depths for quantitative comparison on real colonoscopy data, we evaluate our method qualitatively by comparing the depth prediction results with other methods. First, we compare our method with Rau *et al.* [11] and show some examples in Fig. 4. For the first three examples, we observe the wrongly predicted location of the lumen, missed polyps, and misinterpreted geometry of the lumen respectively in the results generated by Rau *et al.*. For the last three examples, we can see that our method generates more accurate predictions, proving that our model better captures geometric structure details.

We also verify our model in regards to the temporal consistency of the depth estimation. As shown in Fig. 5, without supervision by temporal consistency, the baseline model tends to predict discontinuous depths on the polyp surface due

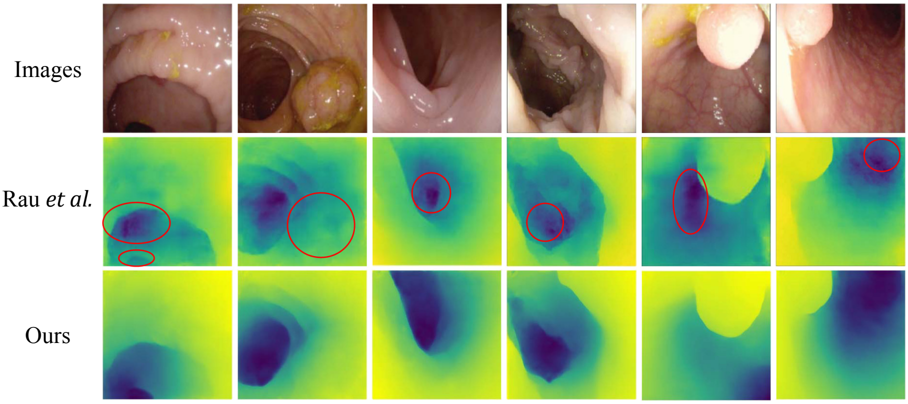

Fig. 4. Comparison of our method with Rau *et al.* [11]. The red ellipses highlight the inaccurate depth predictions such as wrong locations of the lumen, missed polyp, and misinterpreted geometry of the lumen.

to the specular reflection in the colonoscopy frames. These depth noises also lead to the discontinuity between neighboring frames. In comparison, the depths predicted by our fine-tuned model are more spatially smooth and temporally consistent, avoiding the interruption by specular reflections and textures.

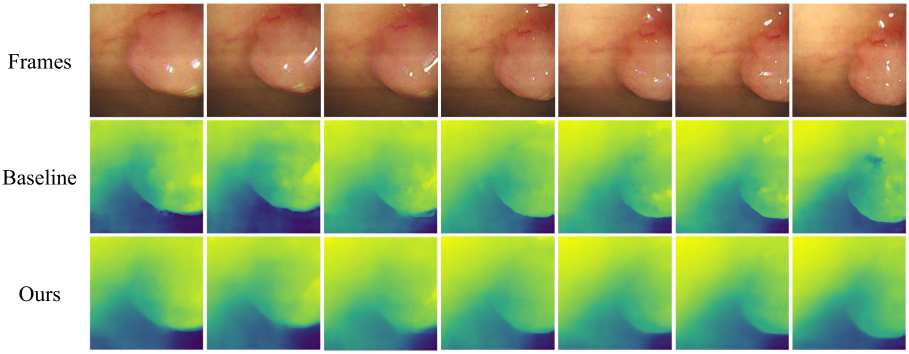

Fig. 5. Depth estimation for adjacent frames in a real colonoscopy video. Compared with the results generated by the baseline model, our model produces more consistent results avoiding the noises caused by specular reflections and textures.

4 Conclusion

We propose a novel depth estimation approach for colonoscopy images that makes full use of both synthetic and real data. Considering the depth estimation task as an image translation problem, we employ a conditional generative network as the backbone model. While the synthetic dataset which contains image-and-depth pairs provides precise supervision on the depth estimation network, we exploit unlabeled real colonoscopy videos for self-supervision. We designed a masked gradient warping loss to ensure the temporal consistency of the estimated depth maps of two neighboring frames during network training. The experimental results demonstrate that our method produces more accurate and temporally consistent depth estimation for both synthetic and real colonoscopy videos. The robust depth estimation will facilitate the accuracy of many downstream medical analysis tasks, such as polyp diagnosis and 3D reconstruction, and assist colonoscopists in polyp localization and removal in the future.

Acknowledgments. We acknowledge funding from National Natural Science Foundation of China under Grants 61976007 and 62076230.

References

1. Arnold, M., Sierra, M.S., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global patterns and trends in colorectal cancer incidence and mortality. Gut **66**(4), 683–691 (2017). https://doi.org/10.1136/gutjnl-2015-310912
2. Freedman, D., Blau, Y., Katzir, L., Aides, A., Shimshoni, I., Veikherman, D., Golany, T., Gordon, A., Corrado, G., Matias, Y., Rivlin, E.: Detecting deficient coverage in colonoscopies. IEEE Trans. Med. Imag. **39**(11), 3451–3462 (2020). https://doi.org/10.1109/TMI.2020.2994221
3. Hong, D., Tavanapong, W., Wong, J., Oh, J., de Groen, P.C.: 3D reconstruction of virtual colon structures from colonoscopy images. Comput. Med. Imag. Graph. **38**(1), 22–33 (2014). https://doi.org/10.1016/j.compmedimag.2013.10.005
4. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632
5. Itoh, H., et al.: Towards automated colonoscopy diagnosis: Binary polyp size estimation via unsupervised depth learning. In: Medical Image Computing and Computer Assisted (MICCAI 2018), pp. 611–619 (2018)
6. Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Trans. Med. Imag. **39**(5), 1438–1447 (2020). https://doi.org/10.1109/TMI.2019.2950936
7. Ma, R., Wang, R., Pizer, S., Rosenman, J., McGill, S.K., Frahm, J.M.: Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions. In: Medical Image Computing and Computer Assisted Intervention, pp. 573–582 (2019). https://doi.org/10.1007/978-3-030-32254-0_64
8. 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). https://doi.org/10.1016/j.media.2018.06.005
9. Nadeem, S., Kaufman, A.: Depth reconstruction and computer-aided polyp detection in optical colonoscopy video frames. arXiv preprint arXiv:1609.01329 (2016)
10. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill (2016). https://doi.org/10.23915/distill.00003

11. Rau, A., Edwards, P.E., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D.: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int. J. Comput. Assist. Radiol. Surg. **14**(7), 1167–1176 (2019). https://doi.org/10.1007/s11548-019-01962-w
12. Schonberger, J.L., Frahm, J.: Structure-from-motion revisited. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016). https://doi.org/10.1109/CVPR.2016.445
13. Stark, U.A., Frese, T., Unverzagt, S., Bauer, A.: What is the effectiveness of various invitation methods to a colonoscopy in the early detection and prevention of colorectal cancer? protocol of a systematic review. Syst. Rev. **9**(1), 1–7 (2020). https://doi.org/10.1186/s13643-020-01312-x
14. Sun, D., Yang, X., Liu, M., Kautz, J.: PWC-net: CNNs for optical flow using pyramid, warping, and cost volume. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8934–8943 (2018). https://doi.org/10.1109/CVPR.2018.00931
15. Waluga, M., Zorniak, M., Fichna, J., Kukla, M., Hartleb, M.: Pharmacological and dietary factors in prevention of colorectal cancer. J. Physiol. Pharmacol. **69**(3) (2018). https://doi.org/10.26402/jpp.2018.3.02
16. Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018). https://doi.org/10.1109/CVPR.2018.00917
17. Widya, A.R., Monno, Y., Okutomi, M., Suzuki, S., Gotoda, T., Miki, K.: Stomach 3D reconstruction based on virtual chromoendoscopic image generation. In: The 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), pp. 1848–1852 (2020). https://doi.org/10.1109/EMBC44109.2020.9176016
18. 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. Trans. Eng. Health Med. **7**, 1–10 (2019). https://doi.org/10.1109/JTEHM.2019.2946802
19. Widya, A.R., Monno, Y., Okutomi, M., Suzuki, S., Gotoda, T., Miki, K.: Self-supervised monocular depth estimation in gastroendoscopy using GAN-augmented images. In: Medical Imaging 2021: Image Processing. vol. 11596, p. 1159616 (2021). https://doi.org/10.1117/12.2579317
20. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE J. Biomed. Health Inform. **21**(1), 65–75 (2017). https://doi.org/10.1109/JBHI.2016.2637004
21. Zhang, R., Zheng, Y., Poon, C.C., Shen, D., Lau, J.Y.: Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Patt. Recogn. **83**, 209–219 (2018). https://doi.org/10.1016/j.patcog.2018.05.026
22. Zhao, Q., Price, T., Pizer, S., Niethammer, M., Alterovitz, R., Rosenman, J.: The endoscopogram: A 3D model reconstructed from endoscopic video frames. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI 2016), pp. 439–447 (2016). https://doi.org/10.1007/978-3-319-46720-7_51
23. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6612–6619 (2017). https://doi.org/10.1109/CVPR.2017.700