Structure-preserving Image Translation for Depth Estimation in Colonoscopy

Shuxian Wang[0000−0002−0184−3021], Akshay Paruchuri[0000−0003−4664−3186], Zhaoxi Zhang[0009−0007−7193−1426], Sarah McGill[0000−0002−4006−2703], and Roni Sengupta[0000−0001−5914−3469]

University of North Carolina at Chapel Hill, Chapel Hill NC 27514, USA

Abstract. Monocular depth estimation in colonoscopy video aims to overcome the unusual lighting properties of the colonoscopic environment. One of the major challenges in this area is the domain gap between annotated but unrealistic synthetic data and unannotated but realistic clinical data. Previous attempts to bridge this domain gap directly target the depth estimation task itself. We propose a general pipeline of structure-preserving synthetic-to-real (sim2real) image translation (producing a modified version of the input image) to retain depth geometry through the translation process. This allows us to generate large quantities of realistic-looking synthetic images for supervised depth estimation with improved generalization to the clinical domain. We also propose a dataset of hand-picked sequences from clinical colonoscopies to improve the image translation process. We demonstrate the simultaneous realism of the translated images and preservation of depth maps via the performance of downstream depth estimation on various datasets.

Keywords: Image-to-image translation · Depth estimation · Colonoscopy

1 Introduction

Colorectal cancer (CRC) is one of the leading causes of cancer mortality in the United States; the American Cancer Society estimates that there will be over 150,000 new cases and 50,000 deaths in 2024. Increased screening is one of the factors contributing to reductions in mortality [13]. Optical colonoscopy is the gold standard method for CRC screening but its effectiveness is highly dependent upon the skill of the physician performing the examination [9]. Around 20% of potentially pre-cancerous polyps are missed during colonoscopies [12][14].

3D reconstruction from optical colonoscopy video can improve efficacy via guidance and visualization to the physician, automatic measurements, and autonomous navigation. One of the major challenges in this area is the lack of realistic data suitable for training neural networks to perform depth and pose estimation. While synthetic [10] and phantom [2] datasets exist, they do not accurately represent the reflectance properties of *in vivo* tissue. Previous approaches towards closing the domain gap [8][11][15] do not target challenging

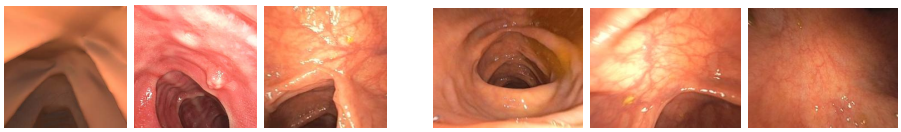
viewpoints making up the majority of colonoscopy videos. In this work, we propose an image translation method that generates realistic-looking video frames from synthetic colonoscopies while preserving the depth information and without requiring complex modeling of mucus and *in vivo* tissue. In this way, we are able to bridge the gap between unrealistic synthetic data with dense ground truth depth annotation and realistic but un-annotated clinical data to improve depth estimation on unseen clinical data. In addition, we introduce two new datasets of manually selected frames from clinical colonoscopies representing viewpoints that are particularly challenging for depth estimation and downstream reconstruction. This data both improves the realism of our image translation results and provides a dataset against which to test the quality of depth estimation results. Code is available at github.com/sherry97/struct-preserving-cyclegan and data at endoscopography.web.unc.edu

2 Related Work

Prior datasets targeting reconstruction from colonoscopy come from clinical procedures (EndoMapper [1], Colon10K [7]), fully synthetic procedures (SimCol3D [10], Zhang et al. [18]), or robotic colonoscopy of a silicone phantom model of the colon (C3VD [2]). Clinical data by nature does not have per-frame depth or pose annotations; while synthetic and phantom data have such annotations, the geometry and light reflectance properties of living tissue is challenging to replicate synthetically and therefore the textures present in the synthetic and phantom data are notably different from those observed in clinical practice (Fig. 1). While the use of image translation to bridge the synthetic to clinical domain gap has been addressed previously (Sec. 2.1), we propose a general modular framework particularly targeting depth estimation (Sec. 2.2) on challenging viewpoints. This is the first work that performs structure-preserving image translation from the synthetic to clinical colonoscopy domain without requiring a pre-trained depth estimator or feature extractor in the target clinical domain.

2.1 Domain gap

Using image translation for colonoscopic depth estimation, Rau et al. [11] propose image-to-depth translation to directly estimate depths from images. In con-



(a) Textures from SimCol3D [10] (left), C3VD [2] (center), and proposed oblique dataset (right).

(b) Viewpoint categories in colonoscopy: axial (left, Colon10K [7]), oblique (center), and *en face* (right).

Fig. 1: Sample frames from various datasets.

trast, Mahmood and Durr [8] combine synthetic depth estimation with real-to-synthetic image translation at inference.

For other tasks, Chen et al. [3] propose a structure-preserving image-to-image generative adversarial network (GAN) to improve segmentation using mutual information in the latent encoding. Similarly, Yoon et al. [17] propose using GAN-based dataset augmentation to boost performance.

For general-purpose image translation, many previous works build upon CycleGAN [19] due to the structure preservation implicit in the cyclical architecture. Cheng et al. [4] present a structure-preserving alternative that decomposes style (extracted via a pretrained autoencoder) from structure (extracted via a pretrained monocular depth estimator).

2.2 Depth estimation

In order to demonstrate the effectiveness of our image translation approach, we use performance on monocular depth estimation as the metric for comparison. Wang et al. [15] propose a self-supervised extension of Monodepth2 [6] for the colonoscopy domain with an iterative refinement step. For general depth estimation, modern Transformer-based methods [5][16] demonstrate high-quality depth estimation results on non-medical data but rely on large training datasets.

3 Data

Generally, we can categorize the viewpoint of a single frame as axial, oblique, or *en face* (Fig. 1b). Oblique and *en face* viewpoints can be challenging for depth estimation due to the lack of strong geometric features. However, they make up about 70% of non-obfuscated frames within a colonoscopy video so reliable depth estimation from these views and their subsequent incorporation into reconstruction provides significant additional information about surface geometry over reconstruction from axial views alone.

In this work, we introduce two distinct datasets: the first of oblique views and the second of *en face* views. Both consist of sequences of consecutive frames manually selected from a library of video recordings of full colonoscopy procedures. The datasets have been curated on the basis of the viewpoint of each frame such that a sequence extends as long as each consecutive frame is of the same viewpoint category modulo gaps of up to 30 consecutive frames with excessive obfuscation (e.g. water drops on the lens).

All frames are pre-processed in the same manner. Using computed camera intrinsics and the Matlab undistortFisheyeImage function, we warp fisheye projection into a pinhole projection. We then crop the image to remove the unused image area and resize to 270×216 pixels. The original videos were recorded using CF and PCF series Olympus colonoscopes with a raw image size of 1350×1080. The UNC Office of Human Research Ethics has determined that this work does not constitute human subject research and does not require Internal Review Board approval.

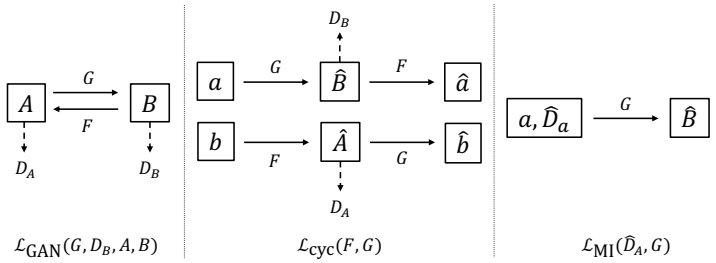


Fig. 2: Image translation framework with image domains A and B, generators $G : A \to B$ and $F : B \to A$, and discriminators D_A and D_B. Let $a \in A, b \in B$ denote data samples and let $\hat{D}_a$ denote the depth map corresponding to sample a. Downstream depth estimation uses output of generator $G(A)$.

Oblique dataset The first dataset, which we call the oblique dataset, consists of sequences manually selected to exclude obfuscated frames, fully axial views, and fully *en face* views. There are 93 sequences totalling 16,756 frames. Each sequence has between 2 and 586 frames, averaging 180 frames per sequence. We randomly divide this dataset into 90% train and 10% test partitions with divisions being made at the sequence (rather than frame) level.

En face dataset The second dataset, which we call the *en face* dataset, consists of sequences manually selected to exclude obfuscated frames and contain only fully *en face* views. There are 14 sequences totalling 816 frames. Each sequence has been 15 and 136 frames, averaging 58 frames per sequence. In this work, we only use this dataset for evaluation due to its small size.

4 Methods

We demonstrate the realism of the image translation result and effectiveness of our proposed structure-preserving loss term via downstream depth estimation. Fig. 2 illustrates our framework. Our image translation result is additionally improved with the use of our proposed data over pre-existing datasets.

Image translation We use the standard CycleGAN losses with generator $G : A \to B$ and discriminator D_B (and similarly generator $F : B \to A$ and discriminator D_A):

$$\mathcal{L}_{\text{GAN}}(G, D_B, A, B) = \mathbb{E}_{b \sim p_{\text{data}}(b)}[\log D_B(b)] + \mathbb{E}_{a \sim p_{\text{data}}(a)}[\log(1 - D_B(G(a)))] \quad (1)$$

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{a \sim p_{\text{data}}(a)}[||F(G(a)) - a||_1] + \mathbb{E}_{b \sim p_{\text{data}}(b)}[||G(F(b)) - b||_1] \quad (2)$$

In order to explicitly constrain the translation to preserve depth information so that the depths and translated image pairs can be used to train supervised depth estimation, we add mutual information loss. Mutual information

circumvents the recursive problem of depth estimation or feature extraction on challenging clinical data.

$$\mathcal{L}_{\text{MI}}(\hat{D}_A, G) = \mathbb{E}_{a\sim p_{\text{data}}(a)} \sum_{i=1}^{|\hat{D}_a|} \sum_{j=1}^{|\hat{I}(G(a))|} \frac{|\hat{D}_i \cap \hat{I}_j|}{N} \log \frac{N|\hat{D}_i \cap \hat{I}_j|}{|\hat{D}_a||\hat{I}(G(a))|} \tag{3}$$

where $\hat{D}_a$ is the ground truth depth map corresponding to the input sample a, $\hat{I}(\cdot)$ is image intensity (average of all color channels), and N is the number of total combinations. We discretize the data into 256 bins for both depths and intensity. This loss is only applied for A $\rightarrow$ B translation. Our full objective is:

$$\begin{aligned}\mathcal{L}(G, F, D_A, D_B) =& \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}(G, D_B, A, B) + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}(F, D_A, B, A) \\ &+ \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G, F) + \lambda_{\text{MI}}\mathcal{L}_{\text{MI}}(\hat{D}_A, G)\end{aligned}$$

We train CycleGAN to perform image translation with SimCol3D as domain A and our proposed oblique dataset as domain B. Table 2 describes the ablations in this portion used for downstream depth estimation.

Depth estimation A significant mismatch between the translated image and the original depth map (lack of structure preservation during translation) will result in poor depth estimation generalization for any model. Here we are interested in depth estimation performance as a metric for the structure preservation through the image translation process and therefore note any architecture could be used. We use the Monodepth2 [6] architecture trained fully supervised from scratch. We pair the RGB result from image translation with the depth map from the original synthetic data for labels. In order to avoid data overlap, we measure performance of all models on C3VD [2]. We convert the fisheye projection to a pinhole projection using the OpenCV undistort function.

Implementation details For image translation, we train the modified CycleGAN using four NVIDIA Titan Xp GPUs for 30 epochs. We use the Adam optimizer and initial learning rate of 2e-4. We use weights $\lambda_{\text{GAN}} = 10.0$, $\lambda_{\text{cyc}} = 0.5$, and $\lambda_{\text{MI}} = 1.0$.

For depth estimation with Monodepth2, we use a Resnet34 backbone and train the model using a NVIDIA Quadro RTX 5000 for 20 epochs with mean squared error loss, Adam optimizer, and initial learning rate of 1e-4. We use data augmentations of random cropping to 256×256 and random horizontal and vertical flipping. At inference, we rescale images to 256×256. All code is implemented using Pytorch.

5 Results

5.1 Image translation

We find that the translation result (Fig. 3) has both improved texture realism and retains the overall geometry of the input image. Most notably, the translation

Table 1: Image translation metrics against oblique dataset. Using $\mathcal{L}_{\text{MI}}$ helps the model produce images more similar to the distribution of test images.

Model	Frechet Inception Distance ↓	Kernel Inception Distance ↓
CycleGAN	2.225	0.220 ± 0.0179
Ours	0.300	0.090 ± 0.0146

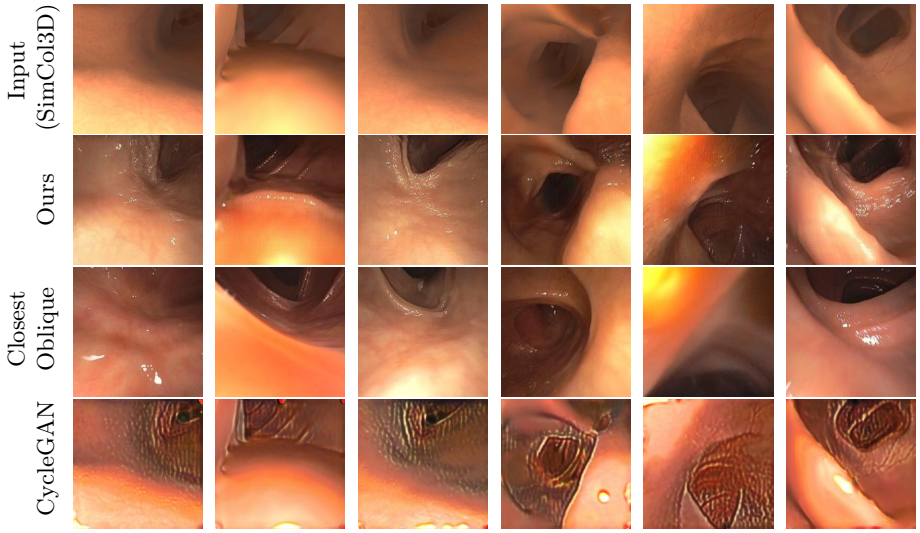


Fig. 3: Examples comparing the SimCol3D input frame, our translation, closest image in oblique dataset via SSIM, and translation with vanilla CycleGAN.

adds the specular points missing from SimCol3D without explicit representation. The specularity is distributed in a manner consistent with our expectation that surfaces closer to the camera and having surface normal directions parallel with the viewing direction will exhibit more specular effects than those either farther from the camera or with normal direction different from that of the viewing direction. In Table 1, we compare translation metrics against translation with $\lambda_{\text{MI}} = 0$ (vanilla CycleGAN) and find the metrics support our perception of improved translation results when $\lambda_{\text{MI}} > 0$.

5.2 Depth estimation

We measure depth estimation performance on C3VD [2] for comparison against baseline models due to its better realism compared to other options but note that the textures and geometries represented in that dataset remain different from those observed in clinical practice. Our qualitative assessment of depth predictions on our proposed oblique and full *en face* datasets demonstrate a notable performance gap on realistic images.

Table 2: Ablations on translation target dataset and use of MI loss. All depth estimations use Monodepth2 architecture, varying in input data.

Depth Estimation Input	**Translation Domain B**	**Uses MI Loss**
Baseline (no translation)	-	-
Ours	oblique	✓
Ours_{CG}	oblique	-
$\text{Ours}_{\text{C10K}}$	Colon10K	✓
$\text{Ours}_{\text{C3VD}}$	C3VD	✓

Table 3: Depth evaluation on C3VD (mm). Best categorical performance highlighted. Multi-shot models train on C3VD while zero-shot rely on generalization. On easy data like C3VD, all experiments perform similarly.

Category	**Model**	**RMSE** ↓	$\textbf{Abs}_{\text{rel}}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Multi-shot	Monodepth2 [5]	18.640	0.297	0.490	0.731	0.861
	UNet [5]	**5.520**	**0.090**	**0.917**	**0.994**	**0.999**
	$\text{Ours}_{\text{C3VD}}$	7.250	0.150	0.794	0.968	0.996
Zero-shot	NormDepth [15]	7.401	**0.169**	0.731	0.948	0.997
	Baseline	9.847	0.205	0.626	0.934	0.991
	$\text{Ours}_{\text{C10K}}$	8.089	0.174	0.735	0.958	0.995
	Ours_{CG}	7.636	0.174	0.730	**0.960**	**0.998**
	Ours	**7.209**	0.174	**0.738**	0.948	0.994

C3VD In Table 3, we provide metrics computed after median rescaling to adjust depth scale across models. For zero-shot models (relying on generalization), we find that our framework produces the best performance in most metrics. We also find that the performance is similar across various models and training datasets. We conclude that the performance on this dataset is satisfactory given the architecture and simplicity of the evaluation dataset, and look for a larger performance gap on more challenging clinical frames.

Oblique In Fig. 4, we show a few examples of depth estimation using NormDepth and our framework evaluated on images from the proposed oblique test partition (additional examples in Fig. S.6). We have not used masking to prevent depth distortions at specular points. Overall, we see that NormDepth is biased towards predicting a depth depression near the center of the frame and poor predictions near occlusion boundaries. Meanwhile, the baseline model produces significant and repeated errors in the depth map at specular points. Our proposed model produces the best representation of rounded haustral ridges and better distinction between structures. Compared to $\text{Ours}_{\text{C10K}}$ and $\text{Ours}_{\text{C3VD}}$, our model produces depths with stronger discontinuities at occlusion boundaries and overall captures a more nuanced and accurate surface geometry.

En face In Fig. 5, we show a few examples of depth estimation using NormDepth and our framework on images from the proposed *en face* dataset (additional examples in Fig. S.7). In these examples, the bias of NormDepth towards predicting

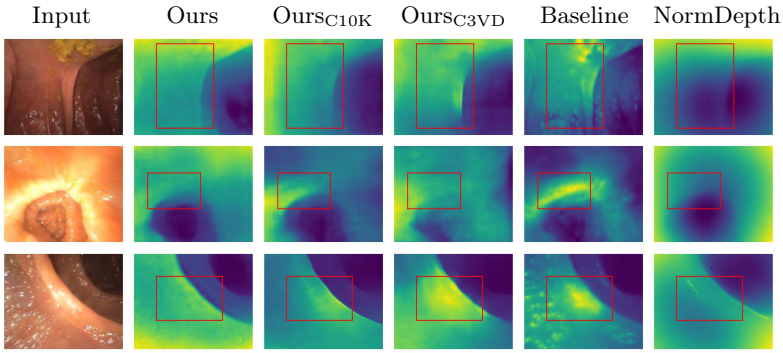


Fig. 4: Depth estimation on oblique dataset. Boxes highlight differences. Image translation framework improves monocular depth estimation in general, with best performance using our proposed dataset as translation target.

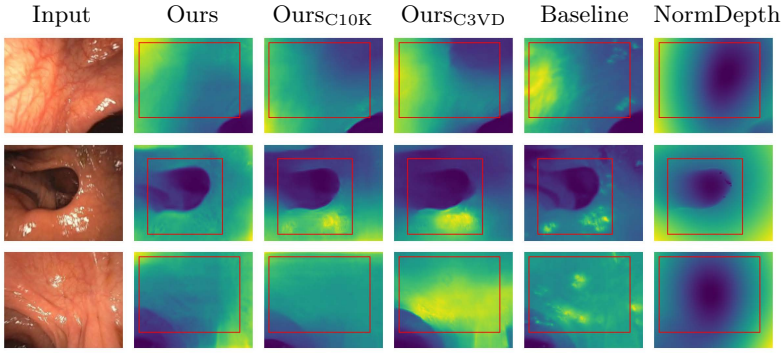


Fig. 5: Depth estimation on *en face* dataset. Boxes highlight differences. Notable improvements from our framework on frames with few geometric features.

a center depth depression is particularly evident, as are the failures of the baseline model in specular areas. Due to the nature of this dataset, there is greater representation of surfaces with strong visual texture from vasculature. Thus we can see that our proposed method has overall improved representation of the overall surface geometry compared to ablations but can also produce distortions to the depth map at regions with strong vascular texture.

6 Conclusions

We have demonstrated that structure-preserving sim2real image translation improves monocular depth estimation in challenging colonoscopic frames. To aid this task, we introduce two datasets of hand-picked sequences from clinical data focusing on viewpoints that are under-represented in existing datasets. The image translation results improve texture realism (especially for specular points) while retaining sufficient depth geometry for successful subsequent training of

depth estimator networks. We provide evaluation of depth estimation on C3VD and qualitative evaluations on our proposed datasets, finding significant performance improvements on challenging frames using this framework.

6.1 Limitations and Future Work

Depth distortions in areas with strongly visible vasculature and few geometric features could be ameliorated by incorporating additional data into the translation target. Future work could focus on applying this approach to pose estimation or other (non-Monodepth2) depth estimation architectures.

Acknowledgments. We thank Stephen Pizer, Sam Ehrenstein, and Julian Rosenman for helpful discussions. Research funding was provided by Olympus Corporation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azagra, P., Sostres, C., Ferrandez, A., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gómez-Rodríguez, J.J., Elvira, R., López, J., Oriol, C., Civera, J., Tardós, J.D., Murillo, A.C., Lanas, A., Montiel, J.M.M.: Endomapper dataset of complete calibrated endoscopy procedures. Scientific Data **10**(1) (October 2023). https://doi.org/10.1038/s41597-023-02564-7, http://dx.doi.org/10.1038/s41597-023-02564-7
2. Bobrow, T.L., Golhar, M., Vijayan, R., Akshintala, V.S., Garcia, J.R., Durr, N.J.: Colonoscopy 3D video dataset with paired depth from 2D-3D registration. Medical Image Analysis p. 102956 (2023)
3. Chen, J., Zhang, Z., Xie, X., Li, Y., Xu, T., Ma, K., Zheng, Y.: Beyond mutual information: Generative adversarial network for domain adaptation using information bottleneck constraint. IEEE Transactions on Medical Imaging **41**(3), 595–607 (2022). https://doi.org/10.1109/TMI.2021.3117996
4. Cheng, M.M., Liu, X.C., Wang, J., Lu, S.P., Lai, Y.K., Rosin, P.L.: Structure-Preserving Neural Style Transfer. IEEE Transactions on Image Processing **29**, 909–920 (2020). https://doi.org/10.1109/TIP.2019.2936746, https://ieeexplore.ieee.org/document/8816670/
5. Eftekhar, A., Sax, A., Bachmann, R., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans (2021)
6. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction (October 2019)
7. Ma, R., McGill, S.K., Wang, R., Rosenman, J., Frahm, J.M., Zhang, Y., Pizer, S.: Colon10k: A benchmark for place recognition in colonoscopy. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1279–1283 (2021). https://doi.org/10.1109/ISBI48211.2021.9433780
8. Mahmood, F., Durr, N.J.: Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Medical Image Analysis **48**, 230–243 (2018). https://doi.org/https://doi.org/10.1016/j.media.2018.06.005, https://www.sciencedirect.com/science/article/pii/S1361841518303761

9. Nierengarten, M.B.: Colonoscopy remains the gold standard for screening despite recent tarnish. Cancer **129** (2023). https://doi.org/10.1002/cncr.34622
10. Rau, A., Bano, S., Jin, Y., Stoyanov, D.: Simcol3D - 3D Reconstruction during Colonoscopy Challenge Dataset (September 2023). https://doi.org/10.5522/04/24077763.v1
11. Rau, A., Edwards, P.E., Ahmad, O.F., Riordan, P., Janatka, M., Lova, L.B., Danail, S.: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. International Journal of Computer Assisted Radiology and Surgery **14**, 1167–1176 (April 2019). https://doi.org/10.1007/s11548-019-01962-w
12. van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., van Deventer, S.J., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: a systematic review. The American Journal of Gastroenterology (2006). https://doi.org/10.1111/j.1572-0241.2006.00390.x
13. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics. CA: A Cancer Journal for Clinicians **74** (2024). https://doi.org/10.3322/caac.21820
14. Vemulapalli, K.C., Lahr, R.E., Rex, D.K.: Most large colorectal polyps missed by gastroenterology fellows at colonoscopy are sessile serrated lesions. Endoscopy International Open (2022). https://doi.org/10.1055/a-1784-0959
15. Wang, S., Zhang, Y., McGill, S.K., Rosenman, J.G., Frahm, J.M., Sengupta, S., Pizer, S.M.: A surface-normal based neural framework for colonoscopy reconstruction. In: Information Processing in Medical Imaging (2023)
16. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv:2401.10891 (2024)
17. Yoon, D., Kong, H.J., Kim, B.S., Cho, W.S., Lee, J.C., Cho, M., Lim, M.H., Yang, S.Y., Lim, S.H., Lee, J., Song, J.H., Chung, G.E., Choi, J.M., Kang, H.Y., Bae, J.H., Kim, S.: Colonoscopic image synthesis with generative adversarial network for enhanced detection of sessile serrated lesions using convolutional neural network. Scientific Reports **12**(1), 261 (January 2022). https://doi.org/10.1038/s41598-021-04247-y, https://www.nature.com/articles/s41598-021-04247-y, number: 1 Publisher: Nature Publishing Group
18. Zhang, S., Zhao, L., Huang, S., Ye, M., Hao, Q.: A template-based 3d reconstruction of colon structures and textures from stereo colonoscopic images. IEEE Transactions on Medical Robotics and Bionics **3**(1), 85–95 (2021). https://doi.org/10.1109/TMRB.2020.3044108
19. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networkss. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)