A Surface-Normal Based Neural Framework for Colonoscopy Reconstruction

Shuxian Wang(✉), Yubo Zhang, Sarah K. McGill, Julian G. Rosenman, Jan-Michael Frahm, Soumyadip Sengupta, and Stephen M. Pizer

University of North Carolina at Chapel Hill, Chapel Hill, USA
{shuxian,zhangyb,jmf,ronisen,pizer}@cs.unc.edu, mcgills@email.unc.edu, rosenmju@med.unc.edu

Abstract. Reconstructing a 3D surface from colonoscopy video is challenging due to illumination and reflectivity variation in the video frame that can cause defective shape predictions. Aiming to overcome this challenge, we utilize the characteristics of surface normal vectors and develop a two-step neural framework that significantly improves the colonoscopy reconstruction quality. The normal-based depth initialization network trained with self-supervised normal consistency loss provides depth map initialization to the normal-depth refinement module, which utilizes the relationship between illumination and surface normals to refine the frame-wise normal and depth predictions recursively. Our framework's depth accuracy performance on phantom colonoscopy data demonstrates the value of exploiting the surface normals in colonoscopy reconstruction, especially on en face views. Due to its low depth error, the prediction result from our framework will require limited post-processing to be clinically applicable for real-time colonoscopy reconstruction.

Keywords: Colonoscopy · 3D reconstruction · surface normal

1 Introduction

Reconstructing the 3D model of colon surfaces concurrently during colonoscopy improves polyp (lesion) detection rate by lowering the percentage of the colon surface that is missed during examination [7]. Often surface regions are missed due to oblique camera orientations or occlusion by the colon folds. By reconstructing the surveyed region, the unsurveyed part can be reported to the physician as holes in the 3D surface (as in Fig. 1). This approach makes it possible to guide the physician back and examine the missing region without delay.

S. Wang and Y. Zhang—These authors contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-34048-2_61.

A. Frangi et al. (Eds.): IPMI 2023, LNCS 13939, pp. 797–809, 2023.
https://doi.org/10.1007/978-3-031-34048-2_61

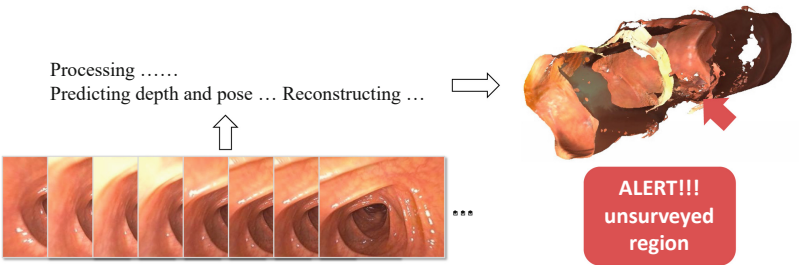

Fig. 1. Reconstructing the 3D mesh from a colonoscopy video in real-time according to the predicted depth and camera pose, allowing holes in the mesh to alert the physician to unsurveyed regions on the colon surface.

To reconstruct colon surfaces from the colonoscopy video, a dense depth map and camera position need to be predicted from each frame. Previous work [12,14] trained deep neural networks to predict the needed information in real time. With the proper help from post-processing [13,15], these methods often are able to reconstruct frames with abundant photometric and geometric features such as in "down-the-barrel" (axial) views where the optical axis is aligned with the organ axis. However, they often fail to reconstruct from frames where the optical axis is perpendicular to the surface ("en face" views). We address the problem of reconstruction from these en face views. In our target colonoscopy application, the geometry of scenes in these two viewpoints are significantly different, manifesting as a difference in depth ranges. In particular, the en face views have near planar geometry, resulting in limited geometric structures informing the photometric cues. As a result, dense depth estimation is challenging using photometric cues alone. However, the characteristics of the endoscopic environment (with a co-located light source and camera located in close proximity to the highly reflective mucus layer coating the colon) mean that illumination is a strong cue for understanding depth and the surface geometry. We capitalize upon this signal to improve reconstruction in en face views. We also aim to yield the reconstruction from frame-wise predictions with minimal post-integration to achieve near real-time execution, which requires strong geometric awareness of the network.

In this work we build a neural framework that fully exploits the surface normal information for colonoscopy reconstruction. Our approach is two-fold, 1) **normal-based depth initialization** (Sect. 3.1) followed by 2) **normal-depth refinement** (Sect. 3.2). Trained with a large amount of clinical data, the normal-based depth initialization network alone can already provide good-quality reconstructions of "down-the-barrel" video segments. To improve the performance on en face views, we introduced the normal-depth refinement module to refine the depth prediction. We find that the incorporation of surface normal-aware losses improves both frame-wise depth estimation and 3D surface reconstruction from the C3DV [3] and clinical datasets, as indicated by both measurements and visualization.

2 Background

Here we describe prior work on 3D reconstruction from endoscopic video, particularly focusing on colonoscopic applications. They usually start with a neural module to provide frame-wise depth and camera pose estimation, followed by an integration step that combines features across a video sequence to generate a 3D surface. With no ground truth from clinical data to supervise the frame-wise estimation network training, some methods transferred the prior learned from synthetic data to real data [4,16,17] while others utilized the self-consistent nature of video frames to conduct unsupervised training [12]. In order to incorporate optimization-based methods to calibrate the results from learning-based methods, Ma et al. [14,15] introduced the system with a SLAM component [5] and a post-averaging step to correct potential camera pose errors; Bae et al. [1] and Liu et al. [13] integrated Structure-from-Motion [18] with the network, trading off time efficiency for better dense depth quality.

When using widely-applied photometric and simple depth consistency objectives [2,25] in training, networks frequently fail to predict high quality and temporally consistent results due to the low geometric texture of endoscopic surfaces and time-varying lighting [23]. The corresponding reconstructions produced by these methods have misaligned or unrealistic shapes as a result. Meanwhile, recent work in computer vision has shown surface normals to be useful for enforcing additional geometric constraints in refining depth predictions [8,20,22] while the relationship between surface normals and scene illumination has been exploited in photometric stereo [9–11,19]. The success of utilizing surface normals in complex scene reconstruction inspires us to explore this property in the endoscopic environment.

3 Methods

Surface normal maps describe the orientation of the 3D surface and reflect local shape knowledge. We incorporate this information in two ways: first, to enhance unsupervised consistency losses in our normal-based depth initialization (Fig. 2a) and second, to allow us to use illumination information in our normal-depth refinement (Fig. 2b). We use this framework as initialization for a SLAM-based pipeline that fuses the frame-wise output into a 3D mesh following Ma et al. [15].

3.1 Normal-Based Depth Initialization

In order to fully utilize the large amount of unlabeled clinical data, our initialization network is trained with self-supervision signals based on the scene's consistency of frames from the same video. We particularly exploit the surface normal consistency in training to deal with the challenges of complicated colon topology in addition to applying the commonly used photometric consistency losses [2,6,25], which are less reliable due to lighting complexity in our application. Trained with the scheme described below, this network produces good depth and camera pose initialization for later reconstruction. We refer to this model as "NormDepth" or "ND" in Sect. 4.

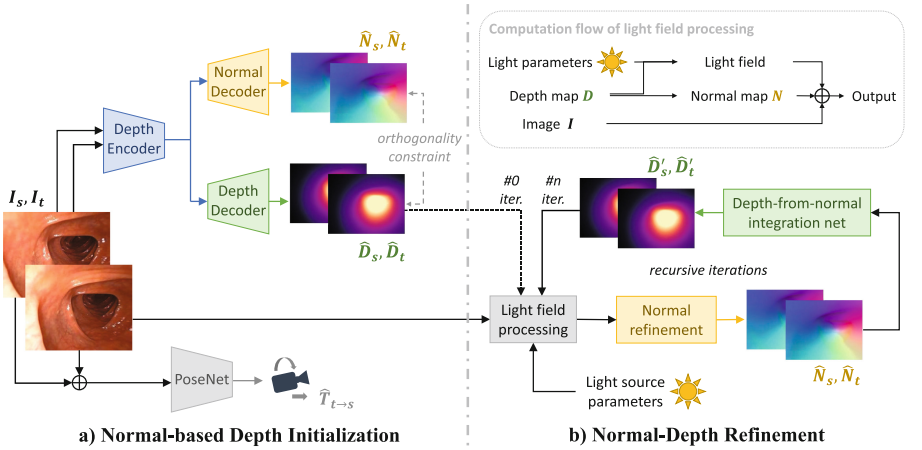

Fig. 2. Our two-fold framework of colonoscopy reconstruction. a) **Normal-based depth initialization** network is trained with self-supervised surface normal consistency loss to produce depth map and camera pose initialization. b) **Normal-depth refinement** framework utilizes the relation between illumination and surface geometry to refine depth and normal predictions.

Background - Projection. The self-supervised training losses discussed in this section are built upon the pinhole camera model and the projection relation between a source view s and a target view t [25]. Given the camera intrinsic K, a pixel p_t in a target view can be projected into the source view according to the predicted depth map $\hat{D}_t$ and the relative camera transformation $\hat{T}_{t\to s}$. This process yields the pixel's homogeneous coordinates $\hat{p}_s$ and its projected depth $\hat{d}_s^t$ in the source view, as in Eq. 1:

$$\hat{p}_s, \hat{d}_s^t \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t \tag{1}$$

Normal Consistency Objective. As the derivative of vertices' 3D positions, surface normals can be sensitive to the error and noise on the predicted surface. Therefore, when the surface normal information is appropriately connected with the network's primary predictions, i.e., the depth and camera pose, utilizing surface normal consistency during training can further correct the predictions and improve the shape consistency.

Let $\hat{N}_t$ be the object's surface normals in the target coordinate system. In the source view's coordinate system, the direction of those vectors depends on the relative camera rotation $\hat{R}_{t\to s}$ (the rotation component of $\hat{T}_{t\to s}$) and should agree with the source view's own normal prediction $\hat{N}_s$; using this correspondence we form the normal consistency objective as

$$L_{norm} = ||\hat{N}_s \langle \hat{p}_s \rangle - \hat{R}_{t \to s} \hat{N}_t||_1 \tag{2}$$

Here, we use the numerical difference between the two vectors (L1 loss) for error. In practice, we find that using angular difference has similar performance.

Surface Normal Prediction. We found that when training with colonoscopy data, computing normals directly from depths as in some previous work [20,21] is less stable and tends to result in unrealistic shapes. Instead, we built the network to output the initial surface normal information individually, and trained it in consensus with depth prediction using L_{orth}:

$$\hat{V}(p) = \hat{D}(p_a) K^{-1} p_a - \hat{D}(p_b) K^{-1} p_b \tag{3}$$

$$L_{orth} = \sum_p \hat{N}(p) \cdot \hat{V}(p) \tag{4}$$

where $\hat{V}(p)$ is the approximate surface vector around p, which is computed from the depths of p_a and p_b, p's nearby pixels. In practice, we apply two pairs of $p_{a/b}$ position combinations, i.e., p's top-left/bottom-right and top-right/bottom-left neighboring pixels. This orthogonality constraint bridges the surface normal and depth outputs so that the geometric consistency constraint on the normal will in turn regularize the depth prediction.

Training Overview. We adapt our depth initialization network from Godard et al. [6] with an additional decoder to produce per-pixel normal vectors besides depths, and apply their implementation of photometric consistency loss L_{photo} and depth smoothness loss L_{sm}. Besides the surface normal consistency, we also enforce the prediction's geometric consistency by minimizing the difference between the predicted depths of the same scene in different frames, as in [2]:

$$L_{depth} = \frac{\left| \hat{D}_s \langle \hat{p}_s \rangle - \hat{D}_s^t \right|}{\hat{D}_s \langle \hat{p}_s \rangle + \hat{D}_s^t} \tag{5}$$

With the per-pixel mask M to mask out the stationary [6], invalid projected or specular pixels, the final training loss to supervise this initialization network is the weighted sum of the above elements, where λ_{1-4} are the hyper-parameters:

$$L^{init} = (L_{photo} + \lambda_1 L_{norm} + \lambda_2 L_{depth}) \odot M + \lambda_3 L_{orth} + \lambda_4 L_{sm} \tag{6}$$

3.2 Normal-Depth Refinement

In the endoscopic environment, there is a strong correlation between the scene illumination from the point light source and the scene geometry characterized by the surface normals. Our normal-depth refinement framework uses a combination of the color image, scene illumination as represented by the light field, and an initial surface normal map as input. We use both supervised and self-supervised consistency losses to simultaneously enforce improved normal map refinement and consistent performance across varying scene illumination scenarios.

Light Field Computation. We use the light field to approximate the amount of light each point on the viewed surface receives from the light source. As in Lichy et al. [10] we parameterize our light source by its position relative to the camera, light direction, and angular attenuation. In the endoscopic environment, the light source and camera are effectively co-located so we take the light source position and light direction to be fixed at the origin O and parallel to the optical axis $\vec{z}$, respectively. Thus for attenuation μ and depth map $\hat{D}$, we define the point-wise light field $\hat{F}$ and the point-wise attenuation $\hat{A}$ as

$$\hat{F} = \frac{O - \hat{D}}{||O - \hat{D}||}, \quad \hat{A} = \frac{(-\sum \hat{F} \cdot \vec{z})^{\mu}}{||O - \hat{D}||^2} \tag{7}$$

For our model input, we concatenate the RGB image, $\hat{F}$, $\hat{A}$, and normal map $\hat{N}$ (computed from the gradient of the depth map) along the channel dimension.

Training Overview. In order to use illumination in colonoscopy reconstruction, we adapt our depth-normal refinement model from Lichy et al. [10] with additional consistency losses and modified initialization. We use repeated iterations for refinement; in order to reduce introduced noise, we use a multi-scale network as in many works in neural photometric stereo [9–11]. After each recursive iteration, we upsample the depth map to compute normal refinement at a higher resolution. We denote n iterations with "$n \times$ NR".

We compute the following losses for each scale, rescaling the ground truth where necessary to match the model output. For the supervised loss L_{gt} for iteration i, we minimize L1 loss on the normal refinement module output $\hat{N}_i$ and the matching ground truth normal map N as well as the L1 loss on the depth-from-normal model output $\hat{D}_i$ and the matching ground truth depth map D. We define a scaling factor $\alpha_i = \frac{\text{median}(D)}{\text{median}(\hat{D}_i)}$.

$$L_{gt} = \sum_i ||N - \hat{N}_i||_1 + ||D - \alpha_i \hat{D}_i||_1 \tag{8}$$

For the depth-from-normal integration module, we compute a normal map $\hat{N}_i'$ from its depth output and minimize L1 loss between it and the input normal map $\hat{N}_i$; this has the effect of imposing the orthogonality constraint between the depth and surface normal maps.

$$L_{dfn} = \sum_i ||\hat{N}_i' - \hat{N}_i||_1 \tag{9}$$

We use a multi-phase training regime for stability. In an iteration, we first train the normal refinement module and substitute an analytical depth-from-normal integration method. For the second phase, we freeze the normal refinement module and train only the depth-from-normal integration module. For the third and final phase of training, we use the normal refinement module and neural integration, optimizing a weighted sum of all losses with hyperparameters λ_1 and λ_2. Thus we define the losses for each phase respectively as follows:

$$L_{refine}^{(1)} = L_{gt} + \lambda_1 L_{norm} \tag{10}$$

$$L_{refine}^{(2)} = L_{dfn} \tag{11}$$

$$L_{refine}^{(3)} = L_{gt} + \lambda_1 L_{norm} + \lambda_2 L_{dfn} \tag{12}$$

4 Experiments

In our experiments, we demonstrate that incorporating surface normal information improves both frame-wise depth estimation and 3D surface reconstruction. We describe the frame-wise depth map improvement over baseline and the effect of various ablations in Sect. 4.1. To evaluate the effect of the frame-wise depth estimation on surface reconstruction, we compare the reconstructions obtained from initializing the SLAM pipeline [15] with the outputs from various methods of frame-wise depth estimation against initialization with ground truth depth maps. We provide a comparison of Chamfer distance [24] on aligned mesh reconstructions in Table 1 and a qualitative comparison in Sect. 4.2. We also provide a qualitative comparison of the surfaces reconstructed from clinical video in Sect. 4.3.

Dataset. To train the normal-based depth initialization network, as well as the self-supervised baseline Monodepth2 [6], we collected videos from 85 clinical procedures and randomly sampled 185k frames as the training set and another 5k for validation. We used the Colonoscopy 3D Video Dataset (C3DV) [3] for the normal-depth refinement module, which provides ground truth depth maps and camera poses from a colonoscopy of a silicone colon phantom. We divide this dataset into 5 randomly-drawn cross-validation partitions with 20 training and 3 testing sequences such that the test sequences do not overlap. The results reported in Sect. 4.1 are the methods' average performance across all folds.

4.1 Frame-Wise Depth Evaluation

We compared our method's depth prediction with several ablations and the baseline against the ground truth in C3DV. Following the practice in Godard

Table 1. Error averaged over 5-fold cross validation test sets of C3DV, ± standard deviation. Best performance in bold. "NormDepth" and "ND" stand for normal-based depth initialization and "$n \times$ NR" stands for normal-depth refinement for n iterations. "NormDepth $-L_{norm}$" denotes NormDepth trained without L_{norm}. "flat init" denotes refinement initialized with planar depth rather than NormDepth output.

Method	Depth Error ↓				Chamfer Distance ↓
	Abs Rel	Sq Rel	RMSE	log RMSE	
Monodepth2 [6]	0.189	2.878	11.779	0.232	0.057 ± 0.039
NormDepth $-L_{norm}$	0.137	1.328	7.411	**0.168**	0.044 ± 0.018
NormDepth	0.141	1.373	7.447	0.173	0.046 ± 0.019
ND init + 1×NR	**0.136**	**1.271**	**7.376**	0.170	**0.038 ± 0.017**
ND init + 4×NR	0.141	1.353	7.479	0.173	0.044 ± 0.018
flat init + 4×NR	0.166	1.927	8.955	0.201	0.047 ± 0.022

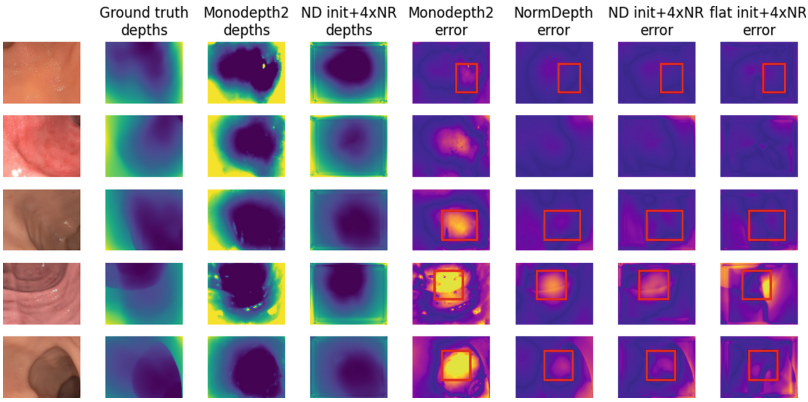

Fig. 3. Example depth predictions and RMSE from C3DV. For the depth maps, darker colors denote more distant depths. For the RMSE, brighter colors denote higher error. Some areas of improvement are highlighted in boxes. (Color figure online)

et al. [6], we rescaled our depth output to match the median of the ground truth and reported 4 pixel-wise aggregated error metrics in Table 1.

Comparing depth prediction errors (Fig. 3), both models using our two-stage method significantly outperform the photometric-based baseline Monodepth2, demonstrating the merit of emphasising geometric features (specifically surface normals) in colonoscopic depth estimation. Meanwhile, although each individual stage of our two-stage method (NormDepth and flat init+NR) already produces relatively good performance, our combined system performs even better and generates the best quantitative result on this dataset (from ND init + 1×NR). Notice that although based on the results from ablation models, the normal consistency loss L_{norm} and multi-iteration of normal refinement quantitatively do not boost performance here due to the nature of C3DV dataset, they are critical for generating better 3D reconstruction shapes (Sects. 4.2 and 4.3).

4.2 C3DV Reconstruction

In this section, we demonstrate the improvement in reconstructions of the C3DV data using our normal-aware methods. In particular, we examine the effects of initializing our SLAM pipeline with the various depth and pose estimation methods. Although C3DV provides a digital model of the phantom, here we compare against the reconstruction produced by using the ground truth depths and poses as initialization to our SLAM pipeline (and refer to this as the ground truth below). In this way, we can control for the impact of the SLAM pipeline in our reconstruction comparison. In Table 1, we measure the Chamfer distance from the ground truth to the reconstructed mesh after ordinary Procrustes alignment and optimizing the scaling factor for Chamfer distance from the ground truth to the reconstruction.

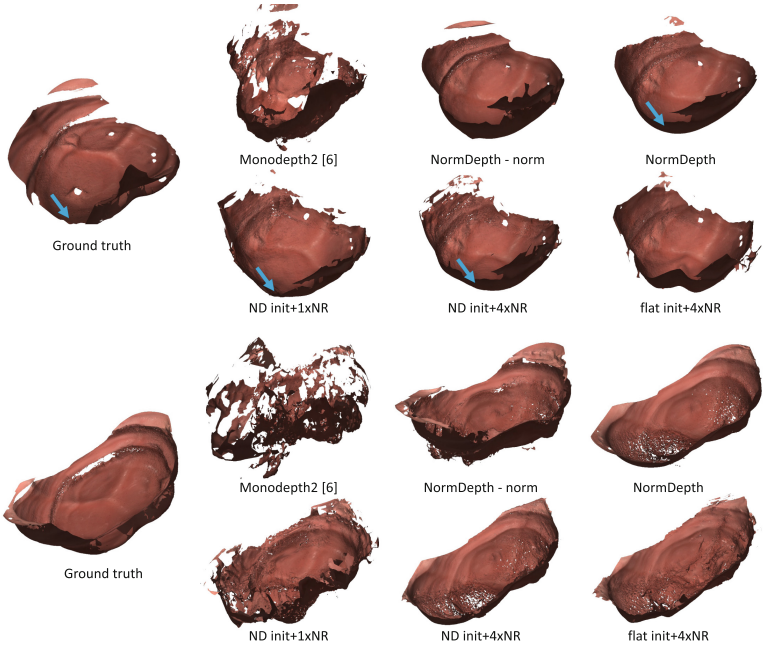

Fig. 4. Example reconstructed sequences from C3DV using various methods of initialization for SLAM pipeline. The more planar shapes observed in the ND init+$n\times$NR compared to NormDepth variations are closer to the ground truth reconstruction while the noisy reconstructions using Monodepth2 and flat init+4$\times$NR are farther from the ground truth. Select areas of improvement highlighted with arrows.

Overall, we find that the performance improvements observed in the frame-wise depth estimation are reflected in the reconstructions as well. Similarly, the weaknesses observed in the frame-wise inference also transfer to the reconstructions. In particular, we note that where significant noise is present in the frame-wise depth estimation for ND init+1xNR and reduced in ND init+4xNR, the corresponding reconstructions reflect the difference in noise as well.

In Fig. 4, we visualize the reconstructions corresponding to example video sequences. In these sequences, we observe that our normal-aware methods significantly outperform the baseline Monodepth2 in qualitative similarity to the ground truth. In addition, we notice that the high curvature of the surface observed in the NormDepth and NormDepth-L_{norm} reconstructions is reduced after refinement, bringing the overall reconstructed result closer to the ground truth.

4.3 Clinical Reconstructions

We tested our trained depth estimation models on clinical colonoscopy sequences and generated 3D reconstruction with the SLAM pipeline. Figure 5 shows the

Fig. 5. 3D reconstruction results on clinical colonoscopy data. Our combined system can handle both "down-the-barrel" and en face views, outperforming the photometric baseline Monodepth2.

reconstructed meshes of two "down-the-barrel" segments (Fig. 5a and b) and an en face segment (Fig. 5c).

The reconstruction quality from the two stages of our method ("ND" and "ND+n×NR") significantly outperforms the photometric baseline Monodepth2. For "down-the-barrel" sequences where features are relatively rich, we expect a generalized cylinder shape with limited sparsity. For sequence (a), we expect two large blind spots due to occlusion by ridges and a slightly curved centerline. For sequence (b), we also expect two large blind spots due to the camera position but fairly dense surface coverage elsewhere. For these sequences, our predictions' shapes are more cylindrical and have surface coverage that more accurately reflect the quantity of surface surveyed compared to the reconstruction produced using Monodepth2. The results also indicate that when trained without the normal consistency loss ($-L_{norm}$), NormDepth tends to predict more artifacts such as the skirt-shape outlier in sequence (a). This demonstrates

the benefit of surface normal information in network training for improved consistency between frames. Meanwhile, using multi-scale iterations of normal-depth refinement can reduce the noise and sparsity of reconstructed meshes compared to a single iteration.

For the en-face sequence (c), we expect a nearly planar surface. Similar to the observations made in reconstructing sequences from C3DV, the high surface curvature produced from the initialization network is reduced after refinement, resulting in a more realistic reconstruction.

5 Conclusion

In this work we introduced the use of surface normal information to improve frame-wise depth and camera pose estimation in colonoscopy video and found that this in turn improves our ability to reconstruct 3D surfaces from videos with low geometric texture. We used a combination of supervised and unsupervised losses to train our multi-stage framework and found significant performance improvements over methods that do not consider surface geometry. We have also shown that the incorporation of normal-aware losses allows us to reconstruct clinical videos of low-texture en face views.

Limitations and Future Work. In this work, we have treated "down-the-barrel" and en face views separately. In practice, colonoscopy videos transition between these two view types, so constructing a framework that can also transition between view types would have significant clinical application; we leave this investigation to future work.

Acknowledgements. We thank Zhen Li and his team at Olympus, Inc. for support and collaboration and Taylor Bobrow for early access to the C3DV dataset.

References

1. Bae, G., Budvytis, I., Yeung, C.-K., Cipolla, R.: Deep multi-view stereo for dense 3D reconstruction from monocular endoscopic video. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 774–783. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_74
2. Bian, J., et al.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. Adv. Neural. Inf. Process. Syst. **32**, 35–45 (2019)
3. 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. arXiv preprint arXiv:2206.08903 (2022)
4. Cheng, K., Ma, Y., Sun, B., Li, Y., Chen, X.: Depth estimation for colonoscopy images with self-supervised learning from videos. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 119–128. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_12
5. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Trans. Pattern Anal. Mach. Intell. **40**(3), 611–625 (2017)

6. 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)
7. Hong, W., Wang, J., Qiu, F., Kaufman, A., Anderson, J.: Colonoscopy simulation. In: Medical Imaging 2007: Physiology, Function, and Structure from Medical Images, vol. 6511, p. 65110R. International Society for Optics and Photonics (2007)
8. Li, B., Huang, Y., Liu, Z., Zou, D., Yu, W.: StructDepth: leveraging the structural regularities for self-supervised indoor depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12663–12673 (2021)
9. Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. In: SIGGRAPH Asia 2018 Technical Papers, p. 269. ACM (2018)
10. Lichy, D., Sengupta, S., Jacobs, D.W.: Fast light-weight near-field photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
11. Lichy, D., Wu, J., Sengupta, S., Jacobs, D.W.: Shape and material capture at home. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
12. Liu, X., et al.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Trans. Med. Imaging **39**(5), 1438–1447 (2019)
13. Liu, X., et al.: Reconstructing sinus anatomy from endoscopic video – towards a radiation-free approach for quantitative longitudinal assessment. In: 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
14. 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: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11768, pp. 573–582. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32254-0_64
15. Ma, R., Wang, R., Zhang, Y., Pizer, S., McGill, S.K., Rosenman, J., Frahm, J.M.: Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy. Med. Image Anal. **72**, 102100 (2021)
16. Mahmood, F., Chen, R., Durr, N.J.: Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans. Med. Imaging **37**(12), 2572–2581 (2018)
17. Mathew, S., Nadeem, S., Kumari, S., Kaufman, A.: Augmenting colonoscopy using extended and directional CycleGAN for lossy image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4696–4705 (2020)
18. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)
19. Xie, W., Nie, Y., Song, Z., Wang, C.C.L.: Mesh-based computation for solving photometric stereo with near point lighting. IEEE Comput. Graphics Appl. **39**(3), 73–85 (2019). https://doi.org/10.1109/MCG.2019.2909360
20. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: LEGO: learning edge with geometry all at once by watching videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 225–234 (2018)
21. Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R.: Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In: Thirty-Second AAAI conference on artificial intelligence (2018)

22. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: MonoSDF: exploring monocular geometric cues for neural implicit surface reconstruction. In: Advances in Neural Information Processing Systems (2022)
23. Zhang, Y., Wang, S., Ma, R., McGill, S.K., Rosenman, J.G., Pizer, S.M.: Lighting enhancement aids reconstruction of colonoscopic surfaces. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 559–570. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78191-0_43
24. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: a modern library for 3D data processing. arXiv:1801.09847 (2018)
25. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1851–1858 (2017)