

# Reconstructing Sinus Anatomy from Endoscopic Video – Towards a Radiation-free Approach for Quantitative Longitudinal Assessment

Xingtong Liu<sup>1</sup>, Maia Stiber<sup>1</sup>, Jindan Huang<sup>1</sup>, Masaru Ishii<sup>2</sup>, Gregory D. Hager<sup>1</sup>, Russell H. Taylor<sup>1</sup>, and Mathias Unberath<sup>1</sup>

<sup>1</sup> The Johns Hopkins University, Baltimore, USA,  
{xingtongliu,unberath}@jhu.edu

<sup>2</sup> Johns Hopkins Medical Institutions, Baltimore, USA

**Abstract.** Reconstructing accurate 3D surface models of sinus anatomy directly from an endoscopic video is a promising avenue for cross-sectional and longitudinal analysis to better understand the relationship between sinus anatomy and surgical outcomes. We present a patient-specific, learning-based method for 3D reconstruction of sinus surface anatomy directly and only from endoscopic videos. We demonstrate the effectiveness and accuracy of our method on *in* and *ex vivo* data where we compare to sparse reconstructions from Structure from Motion, dense reconstruction from COLMAP, and ground truth anatomy from CT. Our textured reconstructions are watertight and enable measurement of clinically relevant parameters in good agreement with CT. The source code is available at <https://github.com/lpp1lpp1920/DenseReconstruction-Pytorch>.

## 1 Introduction

The prospect of reconstructing accurate 3D surface models of sinus anatomy directly from endoscopic videos is exciting in multiple regards. Many diseases are defined by aberrations in human geometry, such as laryngotracheal stenosis, obstructive sleep apnea, and nasal obstruction in the head and neck region. In these diseases, patients suffer significantly due to the narrowing of the airway. While billions of dollars are spent to manage these patients, the outcomes are not exclusively satisfactory. An example: The two most common surgeries for nasal obstruction, septoplasty and turbinate reduction, are generally reported to *on average* significantly improve disease-specific quality of life [1], but evidence suggests that these improvements are short term in more than 40% of cases [2,3]. Some hypotheses attribute the low success rate to anatomical geometry but there are no objective measures to support these claims. The ability to analyze longitudinal geometric data from a large population will potentially help to better understand the relationship between sinus anatomy and surgical outcomes. In current practice, CT is the gold standard for obtaining accurate 3D information about patient anatomy. However, due to its high cost and use of ionizing radiation, CT scanning is not suitable for longitudinal monitoring

of patient anatomy. Endoscopy is routinely performed in outpatient and clinic settings to qualitatively assess treatment effect, and thus, constitutes an ideal modality to collect longitudinal data. In order to use endoscopic video data to analyze and model sinus anatomy in 3D, methods for 3D surface reconstruction that operate solely on endoscopic video are required. The resulting 3D reconstructions must agree with CT and allow for geometric measurement of clinically relevant parameters, e. g. aperture and volume.

**Contributions.** To address these challenges, we propose a patient-specific learning-based method for 3D sinus surface reconstruction from endoscopic videos. Our textured reconstructions are watertight and enable measurement of clinically relevant parameters in good agreement with CT. We extensively demonstrate the effectiveness and accuracy of our method on *in* and *ex vivo* data where we compare to sparse reconstructions from Structure from Motion (SfM), dense reconstruction from COLMAP [4], and ground truth anatomy from CT.

**Related Work.** Many methods to estimate surface reconstruction from endoscopic videos have been proposed. SfM-based methods aim for texture smoothness [5,6,7] and provide a sparse or dense reconstructed point cloud that is then processed by a surface reconstruction method, such as Poisson reconstruction [8]. Unfortunately, there are no guarantees that this approach will result in reasonable surfaces specifically when applied to anatomically complex structures, such as the nasal cavity in Fig. 4. Shape-from-Shading methods are often combined with fusion techniques, such as [9,10,11], and often require careful photometric calibration to ensure accuracy. Reconstruction with tissue deformation are handled in [12,13,14]. In intra-operative scenarios, SLAM-based methods [13,14,15] are preferable as they optimize for near real-time execution. Learning-based methods [15,16] take advantage of deep learning advancements in depth and pose estimation to improve model quality.

## 2 Methods

**Overall Pipeline.** The goal of our proposed pipeline is to automatically reconstruct a watertight textured sinus surface from an unlabeled endoscopic video. The pipeline, shown in Fig. 1, has three main components: 1) SfM based on dense point correspondences produced by a learning-based descriptor; 2) depth estimation; and 3) volumetric depth fusion with surface extraction. SfM identifies corresponding points across the video sequence and uses these correspondences to calculate both, a sparse 3D reconstruction of these points and the camera trajectory generating the video. By replacing local with learning-based descriptors during the point correspondence stage of SfM, we are able to improve the density of sparse reconstruction and the completeness of the estimated camera trajectory. We refer to this process as *Dense Descriptor Extraction*. Reliable and complete reconstructions directly from SfM are important, because these results are subsequently used for two purposes: 1) they provide self-supervisory signals for fine-tuning two learning-based modules, i. e. *Dense Descriptor Extraction* and *Depth Estimation*; 2) they are used in the *Depth Fusion & Surface Extraction*

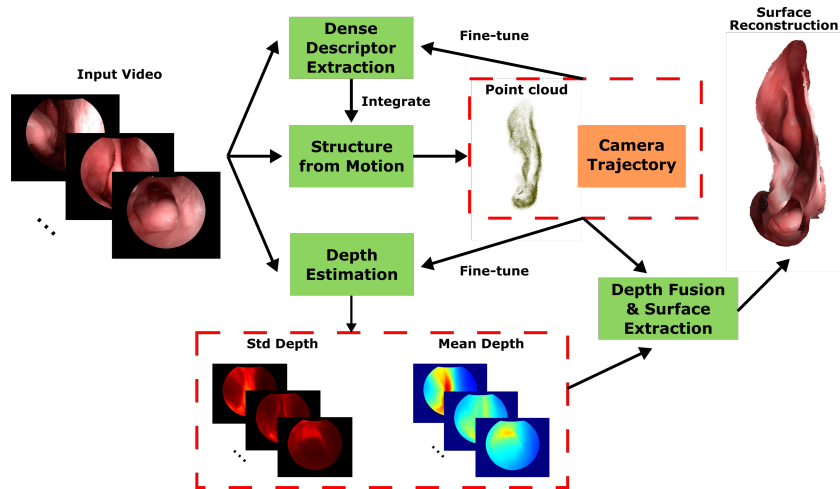


Fig. 1: **Overall pipeline.** Note that part of the surface reconstructions is removed in the figure to display internal structures.

module to guide the fusion procedure. *Depth Estimation* provides dense depth measurements for all pixels in every video frame that are then aggregated over the whole video sequence using *Depth Fusion & Surface Extraction*.

**Training procedure.** The two learning-modules used in our approach, namely *Dense Descriptor Extraction* and *Depth Estimation*, are both self-supervised in that they can be trained on video sequences with corresponding SfM results obtained using a conventional hand-crafted feature descriptor. This training strategy is introduced in [17,18]. Before training the complete pipeline here, we assume that both the aforementioned modules were pre-trained using such a self-supervised strategy. Then, the training order is as follows. First, the pre-trained dense descriptor extraction network is first used to establish correspondences that produce an SfM result. If the result is unsatisfactory, the dense descriptor extraction network will be fine-tuned with this SfM result for bootstrapping. This process can be repeated if necessary. We found the pre-trained dense descriptor extraction network to generalize well to unseen videos so that iterative fine-tuning was not required. Then, the depth estimation network is fine-tuned using the patient-specific dense descriptor extractor and SfM results to achieve the best performance on the input video. Each module in the pipeline is introduced below, please refer to the supplementary material for details of the implementation such as network architecture and loss design.

**Structure from Motion with Dense Descriptor.** SfM simultaneously estimates a camera trajectory and a sparse reconstruction from a video. We choose SfM over other multi-view reconstruction methods such as SLAM because SfM is known to produce more accurate reconstruction at the cost of increased run time. Still, it has been shown in [18] that local feature descriptors have difficulty in dealing with smooth and repetitive textures that commonly occur in

endoscopy. In this work, we adopt the learning-based dense descriptor extraction method in [18] to replace the role of local descriptors for pair-wise feature matching in SfM. Intuitively, such a learning-based approach largely improves the matching performance because the Convolutional Neural Network-based architecture enables global context encoding. In addition, the pixel-wise feature descriptor map generated from the network also enables dense feature matching, which eliminates the reliance on repeatable keypoint detections. With a large number of correct pair-wise point correspondences being found, the density of the sparse reconstruction and the completeness of the camera trajectory estimate are largely improved compared with SfM with local descriptors. For each query location in the source image that is suggested by a keypoint detector, a matching location in the target image is searched for. By comparing the feature descriptor of the query location to the pixel-wise feature descriptor map of the target image, a response map is generated. In order to achieve subpixel matching accuracy, we further apply bicubic interpolation to the response map and use the position with the maximum response as the matching location. A comparison of descriptors and the impact on surface reconstruction is shown in Fig. 2.

**Depth Estimation.** Liu *et al.* [17] proposed a method that can train a depth estimation network in a self-supervised manner with sparse guidance from SfM results and dense inter-frame geometric consistency. In this work, we adopt a similar self-supervision scheme and the network architecture as [17] but assume that depth estimates should be probabilistic because poorly illuminated areas will likely not allow for precise depth estimates. Consequently, we model depth as a pixel-wise independent Gaussian distribution that is represented by a mean depth and its standard deviation. The related training objective is to maximize the joint probability of the training data from SfM given the predicted depth distribution. The probabilistic strategy provides some robustness to outliers from SfM, as shown in Fig. 5a. We also add an appearance consistency loss [19], which is commonly used in self-supervised depth estimation for natural scenes where photometric constancy assumptions are reasonable. This assumption, however, is invalid in endoscopy and cannot be used for additional self-supervision. Interestingly, the pixel-wise descriptor map from *Dense Descriptor Extraction* module is naturally illumination-invariant and provides a dense signal. It can thus be interpreted in analogy to appearance consistency, where appearance is now defined in terms of descriptors rather than raw intensity values. This seemed to further improve the performance, which is qualitatively shown in Fig. 5a. Based on the sparse supervision from SfM together with the dense constraints of geometric and appearance consistency, the network learns to predict accurate dense depth maps with uncertainty estimates for all frames, which are fused to form a surface reconstruction in the next step.

**Depth Fusion and Surface Extraction.** We apply a depth fusion method [20] based on truncated signed distance functions [21] to build a volumetric representation of the sinus surface. Depth measurements are propagated to a 3D volume using ray-casting from the corresponding camera pose and the corresponding uncertainty estimates determine the slope of the truncated signed dis-



tance function for each ray. We used SfM results to re-scale all depth estimates before the fusion to make sure all estimates are scale-consistent. To fuse all information correctly, the camera poses estimated from SfM are used to propagate the corresponding depth estimates and color information to the 3D volume. Finally, the Marching Cubes method [22] is used to extract a watertight triangle mesh surface from the 3D volume.

### 3 Experiments

**Experiment Setup.** The endoscopic videos used in the experiments were acquired from eight consenting patients and five cadavers under an IRB approved protocol. The anatomy captured in the videos is the nasal cavity. The total time duration of videos is around 40 minutes. Because this method is patient-specific, all data are used for training. All processing related to the proposed pipeline used 4-time spatially downsampled videos, which have a resolution of  $256 \times 320$ . SfM was first applied with SIFT [23] to all videos to generate sparse reconstructions and camera trajectories. Results of this initial SfM run were used to pre-train the depth estimation and dense descriptor extraction networks until convergence. Note that the pre-trained depth estimation network was not trained with appearance consistency loss. For evaluation of each individual video sequence, SfM was applied again with the pre-trained dense descriptor extraction network to generate a denser point cloud and a more complete camera trajectory. The depth estimation network, now with appearance consistency loss, were fine-tuned with the updated and sequence-specific SfM results. Note that if the pre-trained descriptor network cannot produce satisfactory SfM results on the new sequence, descriptor network fine-tuning and an extra SfM run with fine-tuned descriptor are required. All experiments were conducted on one NVIDIA TITAN X GPU. The registration algorithm used for evaluation is based on [24] to optimize over similarity transformation.

**Agreement with SfM results.** Because our method is self-supervised and SfM results are used to derive supervisory signals, the discrepancy between the surface and sparse SfM reconstruction should be minimal. To evaluate the consistency between our surface reconstruction and the sparse reconstruction, we calculated the point-to-mesh distance between the two. Because scale ambiguity is intrinsic for monocular-based surface reconstruction methods, we used the CT surface models to recover the actual scale for all individuals where CT data are available. For those that do not have corresponding CT data, we used the average statistics of the population to recover the scale. The evaluation was conducted on 33 videos of 13 individuals. The estimated point-to-mesh distance was  $0.34 (\pm 0.14)$  mm. Examples of the sparse and surface reconstruction overlaid with point-to-mesh distance are shown in Fig. 3.

**Consistency against video variation.** Surface reconstruction methods should be insensitive to variations in video capture, such as camera speed. To evaluate the sensitivity of our method, we randomly sub-sampled frames from the original video to mimic camera speed variation. The pipeline was run for

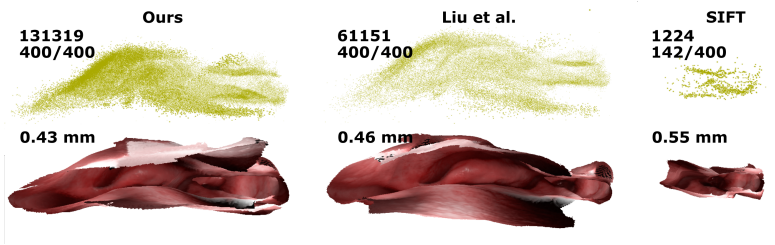


Fig. 2: **Comparison of reconstruction with different descriptors.** We compared the sparse and surface reconstruction using our proposed dense descriptor with those using the descriptor from Liu *et al.* [18] and SIFT [23]. The first row shows sparse reconstructions from SfM using different descriptors. The second row displays surface reconstructions estimated using the proposed method based on the sparse reconstruction. For each column, from top to bottom, the three numbers correspond to the number of points in the sparse reconstruction, number of registered views out of the total ones, and the point-to-mesh distance between the sparse and surface reconstruction. The first row shows that our sparse reconstruction is two times denser than [18]. Surface reconstruction with SIFT covers much less area and has high point-to-mesh distance, which shows the importance of having a dense enough point cloud and complete camera trajectory.

each sub-sampled video and we evaluated the model consistency by aligning surface reconstructions estimated from different subsets. To simulate camera speed variation, out of every 10 consecutive video frames, only 7 frames were randomly selected. We evaluated the model consistency on 3 video sequences that cover the entire nasal cavity of three individuals, respectively. Five reconstructions that were computed from random subsets of each video were used for evaluation. The average residual distance after registration between different surface reconstructions was used as the metric for consistency. The scale recovery method is the same as above. The residual error was  $0.21 (\pm 0.10)$  mm.

**Agreement with dense reconstruction from COLMAP.** We used the ball pivoting [25] method to reconstruct surfaces in COLMAP instead of built-in Poisson [8] and Delaunay [26] methods because these two did not produce reasonable results. Three videos from 3 individuals were used in this evaluation. The qualitative comparison is shown in Fig. 4. The same scale recovery method as above was used. The average residual distance after registration between the surface reconstructions from the proposed pipeline and COLMAP is  $0.24 (\pm 0.08)$  mm. In terms of the runtime performance, given that a pre-trained generalizable descriptor network and depth estimation network exist, our method requires running sparse SfM with a learning-based feature descriptor, fine-tuning depth estimation network, depth fusion, and surface extraction. For the three sequences, the average runtime for the proposed method is 127 minutes, whereas the runtime for COLMAP is 778 minutes.

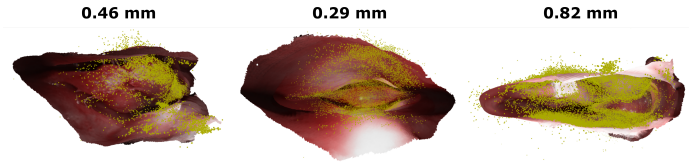


Fig. 3: **Overlay of sparse and surface reconstruction.** Sparse reconstruction from SfM is overlaid with surface reconstruction from the pipeline. The number in each column represents the average point-to-mesh distance.

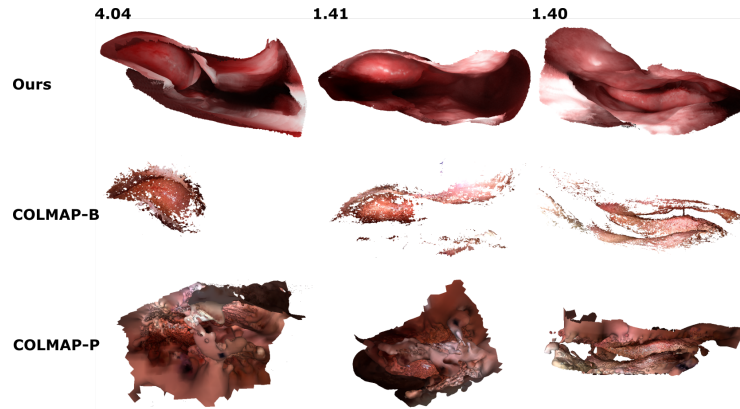


Fig. 4: **Comparison of surface reconstruction from ours and COLMAP.** The number in each column is the ratio of surface area between our reconstruction and COLMAP with ball pivoting [25] (COLMAP-B). Ratios are underestimated because many redundant invalid surfaces are generated in the second row. COLMAP with Poisson [8] (COLMAP-P) is shown in the last row with excessive surfaces removed already.

**Agreement with CT.** Model accuracy was evaluated by comparing surface reconstructions with the corresponding CT models. In this evaluation, two metrics were used: average residual error between the registered surface reconstruction and the CT model, and the average relative difference between the corresponding cross-sectional areas of the CT surface models and the surface reconstructions. The purpose of this evaluation is to determine whether our reconstruction can be used as a low-cost, radiation-free replacement for CT when calculating clinically relevant parameters. To find the corresponding cross-section of two models, the surface reconstruction was first registered to the CT model. The registered camera poses from SfM were then used as the origins and orientations of the cross-sectional planes. The relative differences of all cross-sectional areas along the registered camera trajectory were averaged to obtain the final statistics. This evaluation was conducted on 7 video sequences from 4 individuals. The residual error after registration was  $0.69 (\pm 0.14)$  mm. As a comparison,

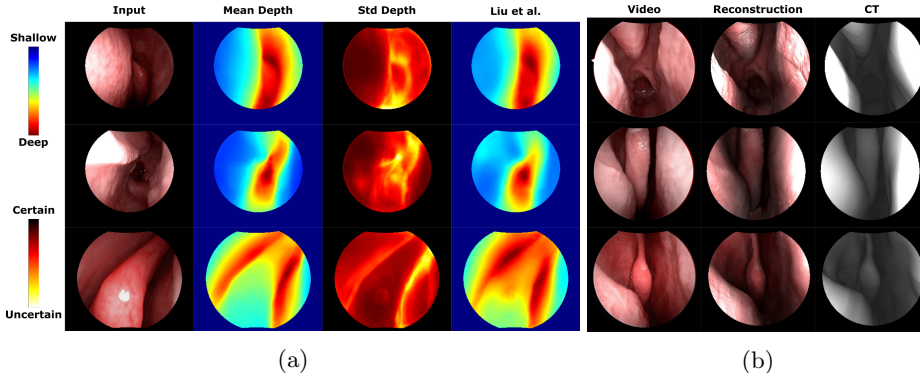


Fig. 5: **(a) Comparison of depth estimation.** By enforcing inter-frame appearance consistency during network training and introducing depth uncertainty estimation, the depth predictions seem to be visually better compared to those from the model trained with the settings in [17]. As can be seen in the third column, higher depth uncertainties were predicted for deeper regions and those where training data is erroneous, such as the specularity in the last row. **(b) Visualization of aligned endoscopic video, dense reconstruction, and CT surface model.** To produce such visualization, our dense reconstruction is first registered to the CT surface to obtain the transformation between the two coordinate systems.

when the sparse reconstructions from SfM are directly registered to CT models, the residual error was  $0.53 (\pm 0.24)$  mm. The smaller error is due to the sparsity and smaller region coverage of the sparse reconstruction compared to ours. In Fig. 5b, a visualization of the video-reconstruction-CT alignment is shown. The cross-sectional surface areas are estimated with an average relative error of  $7 (\pm 2)$  %. This error mainly originates from regions that were not sufficiently visualized during scoping, such as the inferior, middle, and superior meatus. These regions are included in our analysis due to the automation of cross-sectional measurements. In practice, these regions are not commonly inspected as they are hidden beneath the turbinates; if a precise measurement of these areas is desired, small modifications to video capture would allow for improved visualization. Similar to [15], such adjustments can be guided by our surface reconstruction, since the occupancy states in the fusion volume can indicate explicitly what regions were not yet captured with endoscopic video.

## 4 Discussion

**Choice of depth estimation method.** In this work, a monocular depth estimation network is used to learn the complex mapping between the color appearance of a video frame and the corresponding dense depth map. The method in [17] has been shown to generalize well to unseen cases. However, the patient-

specific training in this pipeline may allow for higher variance mappings since it does not need to generalize to other unseen cases. Therefore, a more complex network architecture could potentially further improve the depth estimation accuracy, leading to more accurate surface reconstruction. For example, a self-supervised recurrent neural network that predicts the dense depth map of a video frame based on the current observation and the previous frames in the video could potentially have more expressivity and be able to learn a more complex mapping, such as the method proposed by Wang *et al.* [27].

**Limitations.** First, this pipeline will not work if SfM fails. This could happen in some cases, such as in the presence of fast camera movement, blurry images, or tissue deformation. The latter may potentially be tackled by non-rigid SfM [28]. Second, the pipeline currently does not estimate geometric uncertainty of the surface reconstruction. A volume-based surface uncertainty estimation method may need to be developed for this purpose.

## 5 Conclusion

In this work, we proposed a learning-based surface reconstruction pipeline for endoscopy. Our method operates directly on raw endoscopic videos and produces watertight textured surface models that are in good agreement with anatomy extracted from CT. While this method so far has only been evaluated on videos of the nasal cavity, the proposed modules are generic, self-supervised, and should thus be applicable to other anatomies. Future work includes uncertainty estimation on the reconstructed surface models and prospective acquisition of longitudinal endoscopic video data in the clinic.

## References

1. Bezerra, T.F.P., Stewart, M.G., Fornazieri, M.A., de Mendonca Pilan, R.R., de Rezende Pinna, F., de Melo Padua, F.G., Voegels, R.L.: Quality of life assessment septoplasty in patients with nasal obstruction. *Brazilian journal of otorhinolaryngology* **78**(3) (2012) 57–62
2. Hytönen, M., Blomgren, K., Lilja, M., Mäkitie, A.: How we do it: septoplasties under local anaesthetic are suitable for short stay surgery; the clinical outcomes. *Clinical otolaryngology: official journal of ENT-UK; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial Surgery* **31**(1) (2006) 64–68
3. Hytönen, M.L., Lilja, M., Mäkitie, A.A., Sintonen, H., Roine, R.P.: Does septoplasty enhance the quality of life in patients? *European archives of oto-rhinolaryngology* **269**(12) (2012) 2497–2503
4. Schnberger, J.L., Frahm, J.: Structure-from-motion revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016) 4104–4113
5. Phan, T., Trinh, D., 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). (Sep. 2019) 310–314

6. 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 Journal of Translational Engineering in Health and Medicine* **7** (2019) 1–10
7. Qiu, L., Ren, H.: Endoscope navigation and 3d reconstruction of oral cavity by visual slam with mitigated data scarcity. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2018) 2278–22787
8. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the fourth Eurographics symposium on Geometry processing. Volume 7. (2006)
9. Turan, M., Pilavci, Y.Y., Ganiyusufoglu, I., Araujo, H., Konukoglu, E., Sitti, M.: Sparse-then-dense alignment-based 3d map reconstruction method for endoscopic capsule robots. *Mach. Vision Appl.* **29**(2) (February 2018) 345359
10. Tokgozoglu, H.N., Meisner, E.M., Kazhdan, M., Hager, G.D.: Color-based hybrid reconstruction for endoscopy. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. (June 2012) 8–15
11. Karagyris, A., Bourbakis, N.: Three-dimensional reconstruction of the digestive wall in capsule endoscopy videos using elastic video interpolation. *IEEE Transactions on Medical Imaging* **30**(4) (April 2011) 957–971
12. Zhao, Q., Price, T., Pizer, S., Niethammer, M., Alterovitz, R., Rosenman, J.: The endoscopogram: A 3d model reconstructed from endoscopic video frames. In Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W., eds.: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Cham, Springer International Publishing (2016) 439–447
13. Lamarca, J., Parashar, S., Bartoli, A., Montiel, J.: Defslam: Tracking and mapping of deforming scenes from monocular sequences. *arXiv preprint arXiv:1908.08918* (2019)
14. 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 Robotics and Automation Letters* **3**(4) (2018) 4068–4075
15. 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: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer (2019) 573–582
16. Chen, R.J., Bobrow, T.L., Athey, T., Mahmood, F., Durr, N.J.: Slam endoscopy enhanced by adversarial depth prediction. *arXiv preprint arXiv:1907.00283* (2019)
17. 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 transactions on medical imaging* (2019)
18. Liu, X., Zheng, Y., Killeen, B., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Extremely dense point correspondences using a learned feature descriptor. *arXiv preprint arXiv:2003.00619* (2020)
19. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: *CVPR*. (2017)
20. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. (1996) 303–312
21. Zach, C., Pock, T., Bischof, H.: A globally optimal algorithm for robust tv-l 1 range image integration. In: *ICCV, IEEE* (2007) 1–8

22. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: ACM siggraph computer graphics. Volume 21., ACM (1987) 163–169
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **60**(2) (2004) 91–110
24. Billings, S., Taylor, R.: Generalized iterative most likely oriented-point (g-imlop) registration. *IJCARS* **10**(8) (2015) 1213–1226
25. Bernardini, F., Mittleman, J., Rushmeier, H., Silva, C., Taubin, G.: The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics* **5**(4) (Oct 1999) 349–359
26. Cazals, F., Giesen, J.: Delaunay triangulation based surface reconstruction. In: *Effective computational geometry for curves and surfaces*. Springer (2006) 231–276
27. Wang, R., Pizer, S.M., Frahm, J.M.: Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2019) 5555–5564
28. Khan, I.: Robust sparse and dense nonrigid structure from motion. *IEEE Transactions on Multimedia* **20**(4) (April 2018) 841–850