Texture synthesis for realistic-looking virtual colonoscopy using mask-aware transformer

Anonymous Author(s)
Affiliation
Address
email

Abstract

In virtual colonoscopy, computer vision techniques focus on depth estimation, photometric tracking, and simultaneous localization and mapping (SLAM). To narrow the domain gap between virtual and real colonoscopy data, it is necessary to utilize real-world data or employ realistic-looking virtual dataset. We introduce a texture synthesis and outpainting strategy using the Mask-aware-transformer. The method can generate textures for the inner surface suitable for virtual colonoscopy, including realistic-looking, controllable, and variety of synthesized textures. We generated RGB-D dataset employing the generated virtual colonoscopy, resulting in 9 video recordings. Each sequence was generated from distinct colon models, accumulating a total of 14,120 frames, paired with ground truth depth. Evaluating the generalizability across various datasets, the depth estimation model trained on our dataset exhibited superior transfer performance.

1 Introduction

Nowadays, various computer vision techniques are applied in the field of medical images. In the field of colonoscopy, accurate localization and mapping of colon lesions is the major objectivity of traditional colonoscopy [1]. Virtual reality (VR) attempts to construct simulations containing accurate and high-fidelity texture and organ models. Many studies have been conducted to apply depth estimation in virtual colonoscopy, and subsequently, research is being extended to include photometric tracking and simultaneous localization and mapping (SLAM). [2, 3] evaluated the generalizability of the proposed method using a depth estimation model trained on a virtual dataset, [4] constructed a semantic feature matching reconstruction framework using a Phantom dataset, and [5] proposed a spatial navigation scheme. However, if the colon models are not colored and textured as they are in the real colonoscopy case, the domain gap between virtual and real colonoscopy will lead to a performance drop when transferring to real-world data.

To narrow the gap, it is necessary to utilize real-world data or employ simulation datasets, which apply color and texture to colon models based on actual organs [6]. There are several ways to make a colonoscopy dataset, and one way is to use a computational tomography (CT) colonoscopy simulator [7]. Virtual CT colonoscopy, as a non-invasive technology allowing the creation of colonoscope-like inner views of the human colon, has been used to generate synthetic colonoscopy images. This data-driven approach has some advantages over manual approaches such as using colon phantom, since it can generate various colon cases with little laborious human intervention. Therefore, we release an RGB-D dataset created by texture synthesis, extracted from various anatomical locations in real endoscopy dataset. The proposed dataset leverages 9 video sequences that were registered to generate 14,120 total frames with paired ground truth depth, and 3D models. An overview of this process is shown in Figure 1.

Submitted to the Workshop on Deep Generative Models for Health at NeurIPS 2023. Do not distribute.

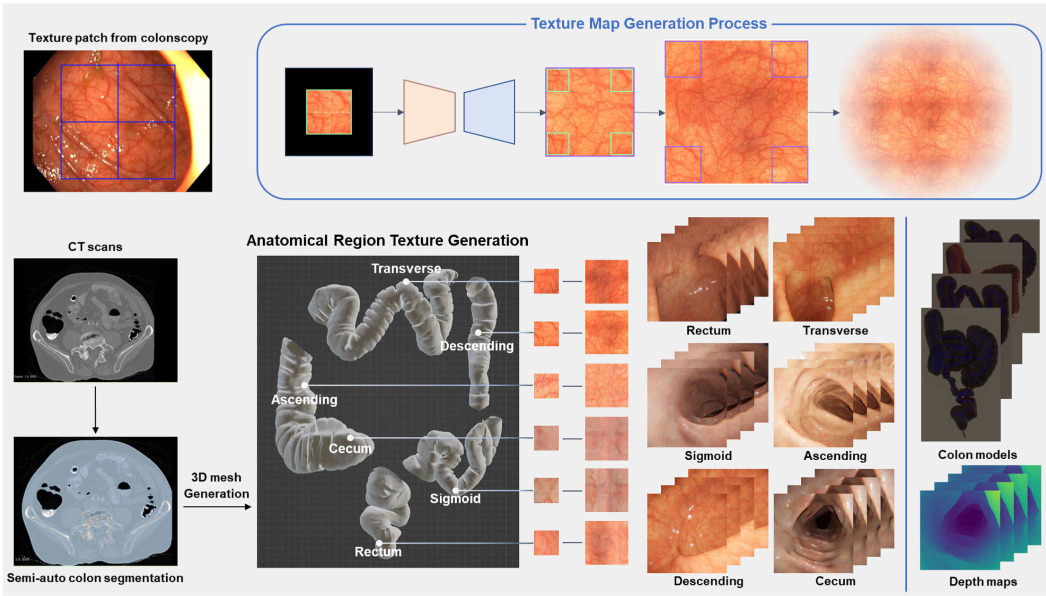

Figure 1: Overview of texture synthesis for realistic-looking virtual colonoscopy using mask-aware-transformer.

2 Method

2.1 Data preparation and preprocessing

Colon texture patch data preparation We made a texture patch dataset extracted from colonoscopy sequences from the Endomapper dataset [8]. The Endomapper dataset contains 96 recordings of endoscopies and provides annotated data, including anatomical regions. For the purpose of maintaining image quality in our patch dataset, we only employed the 'screening' sequences. We extract four texture patches from a single frame by cropping the central part and dividing it into four parts.

Colon mesh generation from CT To visualize the 3D geometry of gastrointestinal organs accurately, we collect computed tomography (CT) images in DICOM format from The Cancer Imaging Archive (TCIA) [9]. We utilized the CT COLONOGRAPHY (ACRIN 6664) collection [10]. For CT mesh generation, we selected 9 patients from the 'no polyp' class. We implemented a medical image segmentation software 3D slicer version 5.2.2, and 3D colons are reconstructed from CT scans.

2.2 Texture synthesis and texture map generation

To achieve realistic-looking texture synthesis, we developed a model based on Mask-aware-transformer (MAT) [11] to generate synthetic colon textures. It consists of a convolutional head designed for tokenization, a transformer body that extracts information through multi-head contextual attention and window shifting. In this study, we trained the model with 16,506 texture patches for 1,000 kimgs with random masking strategy on 4 V100 GPUs and the best FID score was 4.76. In addition, we utilized the texture synthesis model for texture patch cleaning. In many cases, there are unwanted light spots or shaded areas in the texture patches, and in these cases, artifacts may occur in subsequent texture map generation processes. We removed them by masking the unwanted areas and inpainting them using the texture synthesis model.

As shown in Figure 2, we devised an outpainting strategy in our Unity pipeline to generate texture maps that visualize seamlessly, aiming for a realistic-looking virtual colonoscopy. At first, we fixed a seed image at the center and chose to outpaint it with a resolution ranging from 256×256 to 512×512. Based on the synthesized 512×512 image, outpainting was executed in four cardinal directions, yielding a base texture map of 1024×1024 with blank corners. Subsequently, it was rolled to gather the four corners towards the center, followed by a rotation and inpainting process.

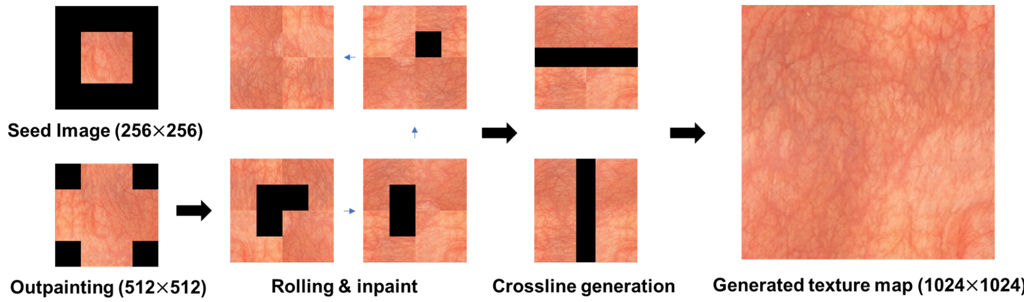

Figure 2: Texture map generation process from single patch image.

After the crossline mask inpainting, the generated texture map ensures a seamless continuation when tiled, eliminating any discordance between the boundaries. As shown in Figure 3, we generated 6 texture maps, each derived from a distinct seed image corresponding to a different anatomical region (rectum, sigmoid, descending, transverse, ascending, and cecum) from real colonoscopy sequences. Final image resolution of the generated texture map of the colon inner surface is 1024×1024.

3 Results

3.1 Visualization in virtual colonoscopy

Generated textures were mapped to the colon model obtained from abdominal CT exam on VR-Caps [12] environment and visualized in virtual colonoscopy manner. The VR-Caps environment is based on the real-time 3D development platform Unity (version 2019.3.3f1) with the integration of Simulation Open Framework Architecture (SOFA) [13]. In addition, the High Definition Render Pipeline (HDRP) has been integrated, which is used to manage lighting conditions while unifying illumination so that all objects in the scene receive and interact with the lighting system. The HDRP shaders offer several options that contribute to a more realistic endoscopic view, such as the light reflection effect, vignetting, and chromatic aberration.

We compared the virtual colonoscopy results of the solid color texture which mimics conventional CT colonoscopy, the texture provided by VR-Caps which is manually made by a professional medical illustrator, and the proposed synthetic texture in Figure 3. In terms of realistic-looking, virtual colonoscopy of synthetic texture showed better visual quality than solid color texture especially with respect to the vascular patterns. The visual quality of VR-Caps is comparable to the proposed synthetic texture, but synthetic texture may be more realistic since the texture is from real colonoscopy images. Another advantage that synthetic texture has is diversity of textures. Additionally, we compared the three texturing methods on virtual colonoscopy images in 6 anatomical locations. As shown in the Figure 4, the solid color texture and VR-Caps texture always represent a frame with the same texture regardless of the anatomical location. Since the proposed texture map could be generated from texture patches for each anatomical location, texture differences according to anatomical locations can be expressed on virtual colonoscopy. This makes it possible to generate more realistic-looking and diverse colonoscopy images.

3.2 Depth estimation

We selected a fully supervised depth estimation model to evaluate the generalizability of the proposed dataset across other datasets. The proposed dataset leverages 9 video sequences, each generated from distinct colon models, accumulating a total of 14,120 RGB frames, paired with ground truth depth. The depth between the far and the near plane is represented by relative values ranging from 0 to 1. We divided the colons into 6, 2, and 1 for the train, validation, and test sets, respectively. Consequently, the synthetic dataset was partitioned into 9,660, 2,650, and 1,810 frames for each of the respective colon models. The results are evaluated across various datasets: a dataset using the VR-Caps texture, a dataset utilizing all 6 synthetic textures, and the average of 6 trained models, each employing one of the synthetic textures respectively. We trained the model with various datasets for 50 epochs. We adopt the standard evaluation metrics: root-mean-squared error (RMSE), the root-mean-squared

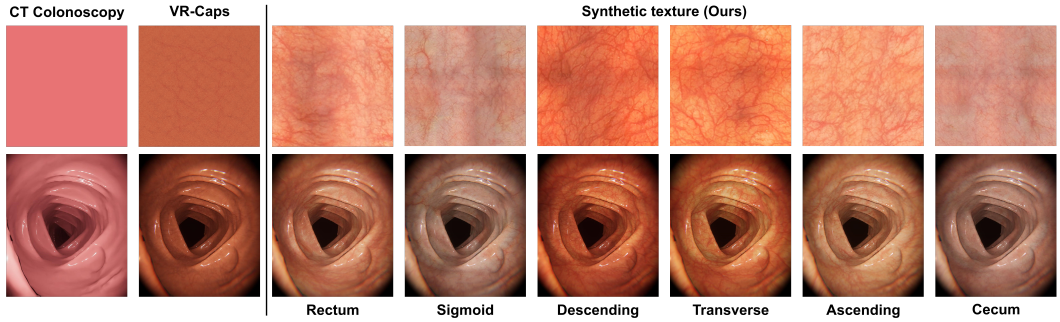

Figure 3: Comparison of texture maps by texturing methods. The first row is texture maps and the second row is virtual colonoscopy scenes.

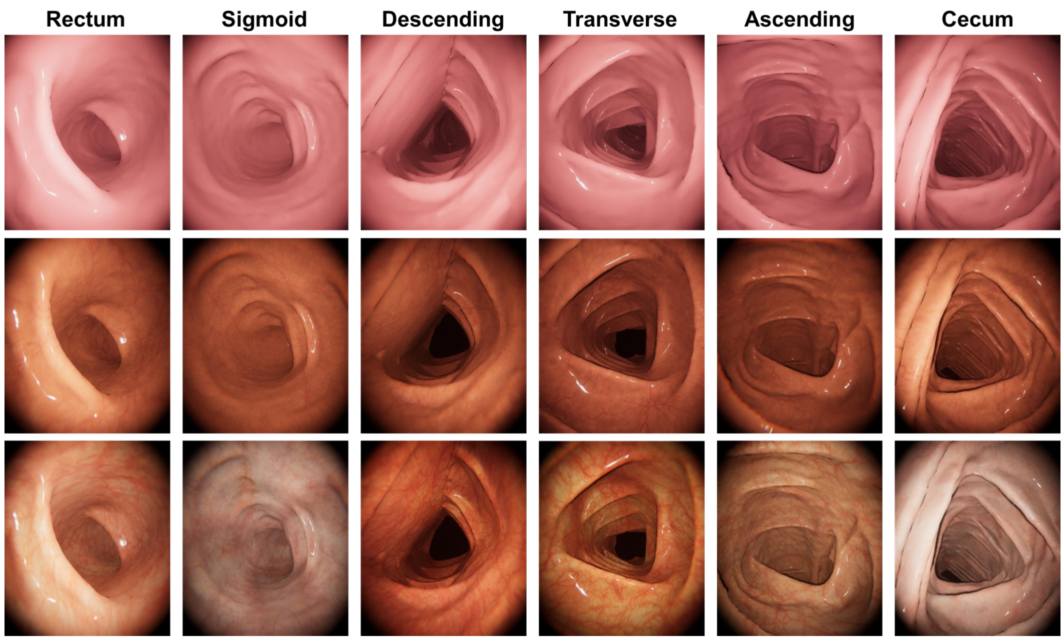

Figure 4: Comparison of virtual colonoscopy scenes by different anatomical regions. The first, second, and third rows represent solid color, VR-caps and synthetic texture, respectively.

logarithmic error (RMSE log), the absolute relative error (Abs Rel), the squared relative error (Sq Rel), and the accuracy (δ<1.25, δ<1.25^2, δ<1.25^3).

ColonDepth dataset [6] The ColonDepth dataset consists of 16,016 RGB images with corresponding depth maps. The data is clustered in groups based on texture and illumination patterns. We resized images to 320 $\times$ 320 and conducted 3-fold validation with 364 T2-L2 frames, and 364 T3-L3 frames.

Scenario dataset [2] The Scenario dataset is crafted to simulate the impact of real-world lighting conditions and currently has 4,500 duodenum frames available. We randomly extracted 750 frames and grouped them into 3 folds, and an evaluation was conducted in a 3-fold validation.

The generalizability of the proposed method is demonstrated across datasets. Given the unique characteristics of T2-L2, models trained with the similar VR-Caps exhibited high accuracy. However, our proposed texture yielded lower errors, as evidenced by achieving RMSE values of 0.124 and Abs Rel values of 0.306. For the T3-L3 dataset, the model trained with a mixed set of all textures exhibited the best performance, verifying the generalizability performance. In addition, on the Scenario dataset, the best performance was observed with the model trained on the mixed-texture dataset, followed

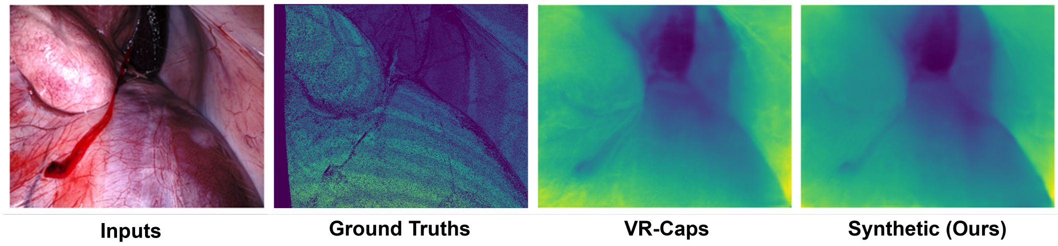

Figure 5: Qualitative results on the real-world images (SCARED dataset [14]).

Table 1: Generalization results of the [6] and the [2]. The best results are in bolded.

Method	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Sq Rel	RMSE	RMSE log
ColonoscopyDepth T2-L2 [6]							
VR-Caps [12]	0.466	**0.773**	**0.935**	0.315	0.054	0.127	**0.333**
Synthetic (All)	0.406	0.715	0.906	0.350	0.066	0.142	0.369
Synthetic (Average)	**0.469**	0.767	0.902	**0.306**	**0.053**	**0.124**	0.344
ColonoscopyDepth T3-L3 [6]							
VR-Caps [12]	0.356	0.597	0.789	0.521	0.127	0.180	0.453
Synthetic (All)	**0.364**	0.601	**0.822**	**0.459**	**0.102**	**0.169**	**0.439**
Synthetic (Average)	0.351	**0.613**	0.812	0.487	0.110	0.172	0.453
Scenario [2]							
VR-Caps [12]	0.249	0.545	0.858	0.468	0.109	0.200	0.452
Synthetic (All)	**0.263**	0.580	**0.898**	0.431	0.093	**0.187**	**0.424**
Synthetic (Average)	0.250	**0.594**	0.881	**0.413**	**0.090**	0.189	0.431

closely by average results of the single-texture trained datasets. As shown in Figure 5, we also compared the result on the SCARED dataset [14]. In contrast to the VR-Caps, where results were noticeably affected by noise like blood and vessels, the proposed method represented more pristine results.

4 Discussion

In this work, we have applied an texture synthesis method to generate texture maps of the colon inner surface learning from real colonoscopy texture patches. To our knowledge, this is the first trial to generate image textures for colon models with the AI-based texture synthesis. This approach has advantages over simple solid color texture mapping used for conventional CT colonoscopy in terms of realistic-looking, especially with respect to the vascular patterns. Also, since texture synthesis is a data-driven approach, it is more efficient than manually made textures used for VR-Caps. In our test environment, it took less than 20 seconds to generate a texture map with a single NVIDIA V100 GPU, and a large number of various texture maps can be easily created from texture patches. The texture mapping method not only enables users to use specific textures they want, but also allows them to select textures according to anatomical locations. In depth estimation experiments, we found that our proposed method exhibited enhanced performance, especially when confronted with diverse textures. A significant advantage of our methodology is its ability to accommodate datasets composed of various textures, enabling customization to fit specific real-world datasets. However, the proposed method has a limitation in considering factors such as light, water, and bubbles, as it primarily focuses on learning from normal mucosal textures. Nonetheless, by implementing a realistic virtual colonoscopy through future studies, it is expected that AI engineers can effectively train and help reduce the gap between virtual and real-world data.

References

[1] Guodao Zhang and Nima Jafari Navimipour. A comprehensive and systematic review of the iot-based medical management systems: Applications, techniques, trends and open issues. *Sustainable Cities and Society*, 82:103914, 2022.

[2] Yongming Yang, Shuwei Shao, Tao Yang, Peng Wang, Zhuo Yang, Chengdong Wu, and Hao Liu. A geometry-aware deep network for depth estimation in monocular endoscopy. *Engineering Applications of Artificial Intelligence*, 122:105989, 2023.

[3] Yuying Liu and Siyang Zuo. Self-supervised monocular depth estimation for gastrointestinal endoscopy. *Computer Methods and Programs in Biomedicine*, page 107619, 2023.

[4] Zhuoyue Yang, Junjun Pan, Ranyang Li, and Hong Qin. Scene-graph-driven semantic feature matching for monocular digestive endoscopy. *Computers in Biology and Medicine*, 146:105616, 2022.

[5] Ameya Pore, Martina Finocchiaro, Diego Dall'Alba, Albert Hernansanz, Gastone Ciuti, Alberto Arezzo, Arianna Menciassi, Alicia Casals, and Paolo Fiorini. Colonoscopy navigation using end-to-end deep visuomotor control: A user study. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9582–9588. IEEE, 2022.

[6] Anita Rau, PJ Eddie Edwards, Omer F Ahmad, Paul Riordan, Mirek Janatka, Laurence B Lovat, and Danail Stoyanov. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *International journal of computer assisted radiology and surgery*, 14:1167–1176, 2019.

[7] Kağan İncetan, Ibrahim Omer Celik, Abdulhamid Obeid, Guliz Irem Gokceler, Kutsev Bengisu Ozyoruk, Yasin Almalioglu, Richard J Chen, Faisal Mahmood, Hunter Gilbert, Nicholas J Durr, et al. Vr-caps: a virtual environment for capsule endoscopy. *Medical image analysis*, 70:101990, 2021.

[8] Pablo Azagra, Carlos Sostres, Ángel Ferrandez, Luis Riazuelo, Clara Tomasini, Oscar León Barbed, Javier Morlana, David Recasens, Victor M Batlle, Juan J Gómez-Rodríguez, et al. Endomapper dataset of complete calibrated endoscopy procedures. *arXiv preprint arXiv:2204.14240*, 2022.

[9] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.

[10] C Daniel Johnson, Mei-Hsiu Chen, Alicia Y Toledano, Jay P Heiken, Abraham Dachman, Mark D Kuo, Christine O Menias, Betina Siewert, Jugesh I Cheema, Richard G Obregon, et al. Accuracy of ct colonography for detection of large adenomas and cancers. *New England Journal of Medicine*, 359(12):1207–1217, 2008.

[11] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[12] Kağan İncetan, Ibrahim Omer Celik, Abdulhamid Obeid, Guliz Irem Gokceler, Kutsev Bengisu Ozyoruk, Yasin Almalioglu, Richard J Chen, Faisal Mahmood, Hunter Gilbert, Nicholas J Durr, et al. Vr-caps: a virtual environment for capsule endoscopy. *Medical image analysis*, 70:101990, 2021.

[13] François Faure, Christian Duriez, Hervé Delingette, Jérémie Allard, Benjamin Gilles, Stéphanie Marchesseau, Hugo Talbot, Hadrien Courtecuisse, Guillaume Bousquet, Igor Peterlik, et al. Sofa: A multi-model framework for interactive physical simulation. *Soft tissue biomechanical modeling for computer assisted surgery*, pages 283–321, 2012.

[14] Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021.