Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2021;10(3):1015-9

Effectiveness of synthetic data generation for capsule endoscopy images

Mehmet Turan

Bogazici University, Institute of Biomedical Engineering , Istanbul, Turkey

Received 19 July 2021; Accepted 22 August 2021
Available online 24.08.2021 with doi: 10.5455/medscience.2021.07.236

Abstract

With advances in digital healthcare technologies, optional therapeutic modules and tasks such as depth estimation, visual localization, active control, automatic navigation, and targeted drug delivery are desirable for the next generation of capsule endoscopy devices to diagnose and treat gastrointestinal diseases. Although deep learning applications promise many advanced functions for capsule endoscopes, some limitations and challenges are encountered during the implementation of data-driven algorithms, with the difficulty of obtaining real endoscopy images and the limited availability of annotated data being the most common problems. In addition, some artefacts in endoscopy images due to lighting conditions, reflections as well as camera view can significantly affect the performance of artificial intelligence methods, making it difficult to develop a robust model. Realistic simulations that generate synthetic data have emerged as a solution to develop data-driven algorithms by addressing these problems. In this study, synthetic data for different organs of the GI tract are generated using a simulation environment to investigate the utility and generalizability of the synthetic data for various medical image analysis tasks using the state-of-the-art Endo-SfMLearner model, and the performance of the models is evaluated with both real and synthetic images. The extensive qualitative and quantitative results demonstrate that the use of synthetic data in training improves the performance of pose and depth estimation and that the model can be accurately generalized to real medical data.

Keywords: Synthetic data generation, capsule endoscopy, depth and pose estimation

Introduction

The use of conventional optical colonoscopy in early diagnosis, prognostic follow-up, and treatment of critical gastrointestinal diseases (GI) is considered the gold standard in the current literature [1,2]. Although colonoscopy has shown clinical efficacy in reducing the incidence of colorectal cancer, accessing the small intestine and obtaining clinically significant images are considered as a challenging procedure due to its complicated anatomical structure [3,4]. Hence, it should be performed by experienced clinicians that requires more time and training. In addition, patients suffer pain and discomfort during this invasive procedure, which can lead to unexpected complications. Wireless capsule endoscopy technology has emerged as a replacement for conventional endoscopes and colonoscopes to eliminate the aforementioned problems and to examine the entire gastrointestinal tract and detect lesions [5-7]. Unlike conventional endoscopes and colonoscopes, capsule endoscopes are pill-shaped, swallowable devices that require minimal preparation before the procedure and allow visualization of the digestive tract by reducing or even eliminating the dose of sedatives that should be administered to the patient [8,9]. In order to extend the scan time of the capsule endoscopy robots and enable active functions such as tissue sampling and therapeutic interventions, sophisticated algorithms are required to be integrated into actively moving capsules [10]. To ensure that these data-driven algorithms perform accurately in clinical cases, the models need to be trained with large amounts of annotated data. However, there are some practical barriers to the creation of such datasets, including privacy, time, lack of expertise for annotation, regulatory concerns related to data collection, and underrepresentation of rare cases. Moreover, some artefacts encountered in real endoscopy images due to the lightning conditions and camera properties can adversely affect the ability of conventional models. Therefore, synthetic data generation and data augmentation methods can be effectively used to improve the performance of artificial intelligence algorithms that are impaired due to limited real endoscopy data [11-13]. To this end, Mahmood and Durr [14] proposed an adversarial training based approach to generate synthetic endoscopy images and proved the success of their approach in monocular depth estimation applications. In [15], Deep Convolutional Generative Adversarial

***Corresponding Author:** Mehmet Turan, Bogazici University, Institute of Biomedical Engineering , Istanbul, Turkey
E-mail: mehmet.turan@boun.edu.tr

Networks (GAN) was utilized to produce synthetic images of malignant mammograms in order to leverage the classification performance of a network in an imbalanced dataset and further analyzes were performed in [16] to determine the quality and realism of the generated images. Although studies on medical image synthesis have been accelerated, especially after the advances in GAN, training neural networks on synthetic data may not be generalizable to real medical images because few models correctly transfer the morphological features of real images to the simulated images. In our previous work [17], a virtual simulation environment was developed using advanced imaging and image processing techniques to generate fully labelled realistic synthetic data consisting of the topology and tissue of the original organ for tasks including disease classification, area coverage and visual odometry. In this new study, a further investigation is presented to explore the applicability and generalizability of synthetic data to neural network performance on both real and synthetic image domains by creating simulated endoscopy images for the organ instances stomach, small intestine and colon of the gastrointestinal tract (GI) using the realistic virtual capsule environment presented in [17].

The rest of this paper is organised as follows: In the materials and methods section, the simulation environment is described and the pipeline for generating synthetic images is presented. In the results section, the efficiency of synthetic data using the state-of-the-art Endo-SfMLearner model for depth and pose estimation use-cases is investigated and the obtained results are presented. Finally, in the discussion and conclusion section, future work is discussed and some concluding remarks are made.

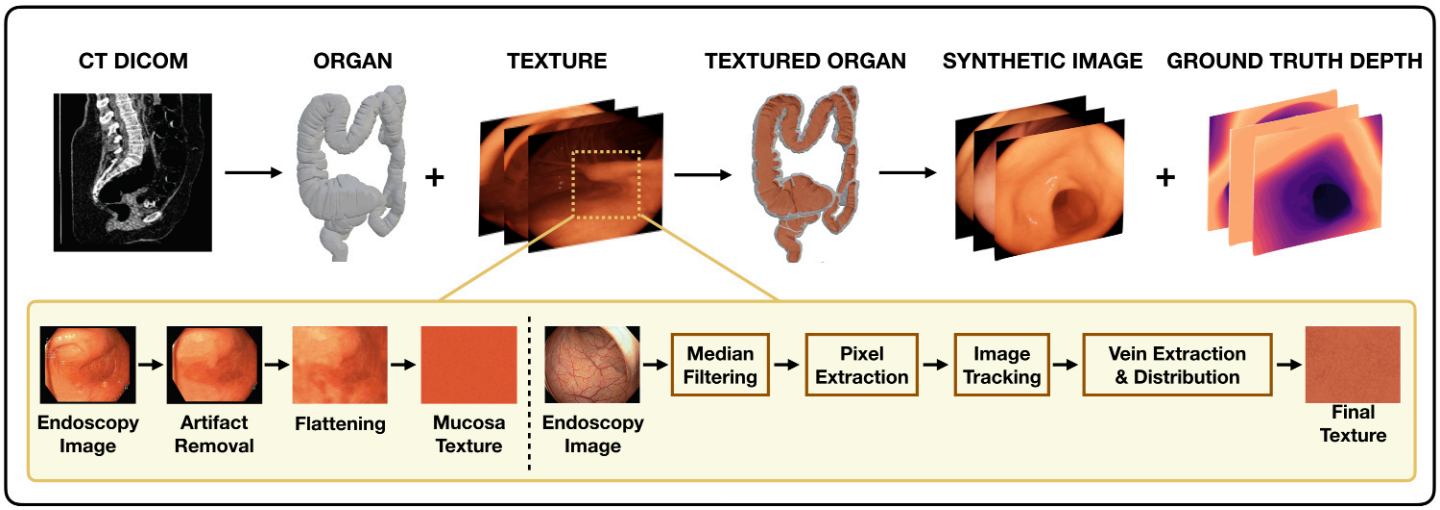


Figure 1. The pipeline for 3D synthetic organ generation and tissue integration is shown schematically. To reconstruct a 3D organ model, computed tomography scans (CT) of patients are used which is followed by the process of removing artifacts and reflections in the endoscopy image to create mucosal tissue layers from real endoscopy images. Then, a uniform color region is selected and this region is flattened to form the main mucosal tissue. For the integration of the vessels into the mucosal tissue, the vessels are extracted from the real endoscopy image using MATLAB and the extracted vessels are added to the tissue image using Gaussian distribution, resulting in the corresponding vascular mesh.

Materials and Methods

Generation of 3-Dimensional Organs in the Simulation Environment

The simulation environment is created using Unity, a real-time 3-dimensional (3D) design and development platform, with the integration of ML -Agent [18] and SOFA [19] as described in [17]. SOFA is an open source software library designed to facilitate the development and testing of medical simulations by enabling the creation of complex medical simulations. ML -Agent, on the other hand, allows Unity to be used as an interactive stage for training intelligent agents with machine learning algorithms. To mimic the mechanics of organ deformation, SofaAPAPI-Unity3D is integrated, an interface that allows Unity PhysX Engines and SOFA to use particularly perfect models for tissue deformation.

Computed tomography images (CT) in DICOM format, openly accessible and anonymized in Cancer Image Archives (TCGA), are used to accurately identify and integrate the 3D geometry of the organs of the gastrointestinal tract in synthetic endoscopy data generation with the simulation environment. Since the archive contains different datasets for different body parts, a set of 46 subjects [20] for the stomach and 825 subjects for the small intestine and colon [21] are identified as reference and included in the study. However, since the dataset is divided into two different subgroups based on the position of the patient (prone and supine), the supine sets are included in the study considering the positions of the patients during capsule endoscopy. An open-source medical imaging application called InVesalius is used to reconstruct 3D organ models from computed tomography scans. The reconstructed 3D model is exported to Blender and MeshLab for processing. To create mucosal overlays, the Kvasir dataset [22] is used, which contains real endoscopy images categorized by different organs of the GI tract. During the generation of main mucosal tissues from this dataset, different endoscopy images are combined in the RGB color scale and placed as tissue overlays on the inner lumen surface of the model to produce faithful mucosal walls without gaps. Vascular networks are then extracted from the endoscopy images and applied to the mucosal tissue images using empirically determined means and standard deviations and a random distribution for vessel size, rotation, and position. In

the final step, the 3D model is reconstructed and the mesh models are decomposed into rectangular segments which are uniformly mirrored onto the UV texture maps generated by the models in a repetitive manner. The pipeline of 3D organ generation is shown in Figure 1.

Synthetic Image Generation

To obtain realistic images from the simulation environment, a cinematic rendering tool is used to mimic the imperfections such as chromatic aberration, visual distortion, specular reflection, and field of view that may occur in real endoscopy images due to lightning conditions and camera characteristics. Three examples of endoscopic videos with pixel-wise depth and pose references at 320x320 pixels, 30fps, containing 2,500 frames of the stomach, 6,700 frames of the small intestine, and 8,200 frames of the colon, are created by moving the monocular camera-enabled virtual capsule through the 3D virtual lumen of the gastrointestinal tract.

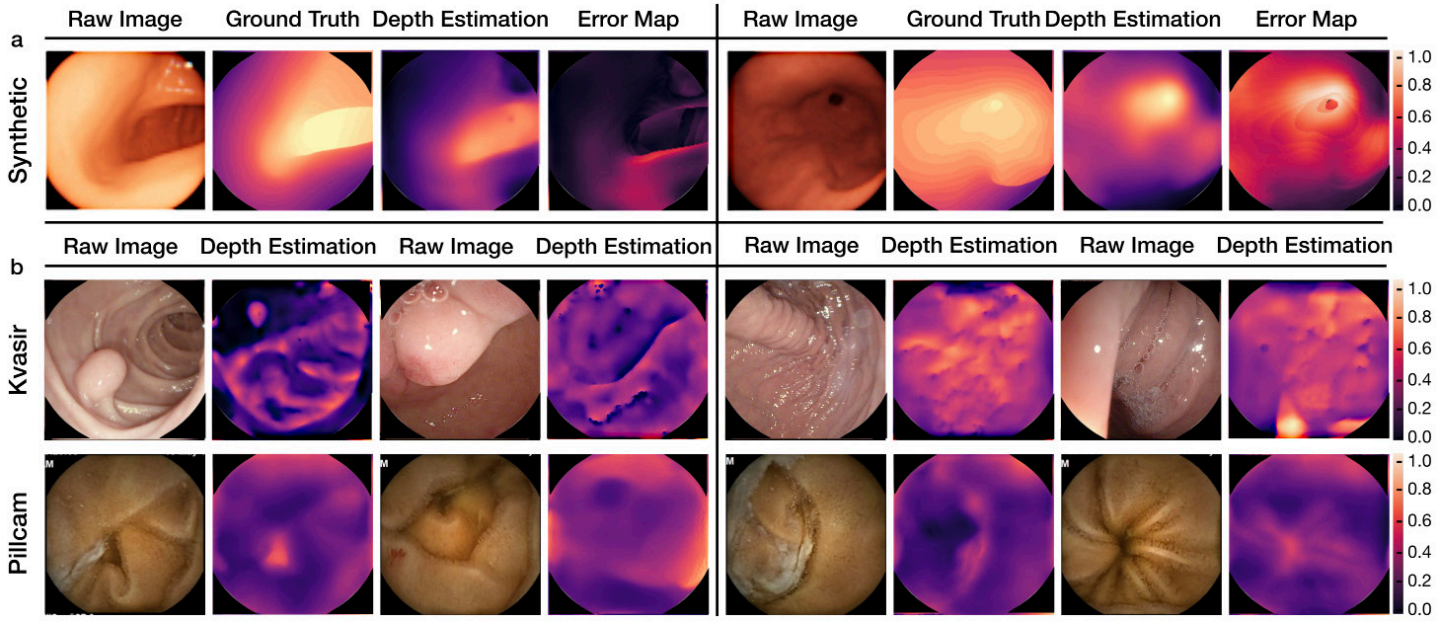


Figure 2. Depth estimation results of the Endo-SfMLearner network trained with synthetic data. a The synthetic raw images generated by the simulation environment are given with the corresponding ground truth depth information, depth estimation and error heatmaps. b For real endoscopy, the Endo-SfMLearner model is tested with the Kvasir and Pillcam datasets for stomach and colon, respectively. Since no reference depth information exists for either dataset, raw images are provided with their depth estimations. The qualitative results illustrate that the model trained with synthetic images gives better estimations for both real and virtual endoscopy data.

Results

To show the efficiency and usefulness of the synthetically generated images, the evaluations are performed in use-cases for depth and pose estimation, and the obtained results are analysed in detail in the following subsections.

Depth Estimation

To investigate the utility of synthetic data on neural network performance in pixel-wise depth estimation, the Endo-SfMLearner algorithm [23] is trained with synthetically generated data only and then tested with both synthetic and real endoscopic images. The training and validation sets consist of 2,400 and 600 images, respectively. The real endoscopy test set consists of images taken in the colon using a Pillcam with a binocular camera and the images from the Kvasir dataset. The neural network is set up with a learning rate of 10-4 and trained for 180 epochs with a batch size of 4 using ADAM optimization. In Figure 2, depth estimation results acquired from each test dataset are illustrated. Since the simulation environment can provide reference depth information for each image generated, error heat maps are also produced in order to demonstrate the predictive success of the model. From the error heatmaps, it can be concluded that the model makes good predictions for pixels representing distant regions, which is confirmed by a decrease in errors in these regions. On the other hand, only estimations are presented for real endoscopy images since their reference depth information is not available. As can be seen in each case, the boundaries and characteristics of areas that differ from the surrounding tissue not only in structure but also in texture are successfully recognized by the model. Overall, the results show that deep learning models trained on synthetic images in a cross-dataset environment exhibit sufficient generalization performance in depth estimation.

Pose Estimation

For a comparative analysis, two different training scenarios are designed on the Endo-SfMLearner model for the pose estimation use-case. In the first case, the artificial neural network is trained on a subset of the EndoSLAM dataset [23] with real porcine samples containing 6-DoF ground truth pose values, with 2,000 input images, a batch size of 4, and a learning rate of 10-4 for 250 epochs. In the second case, the neural network model is fine-tuned with real endoscopy images after pre-training with synthetic data, including 2,000 images generated from the simulation environment, using

the same hyperparameters as in the previous case. Both models are tested with two trajectories of EndoSLAM data containing 1,000 and 900 images of the small intestine and colon, respectively. To assess the performance of the models, three evaluation metrics are used: (1) Absolute Trajectory Error (ATE), which indicates the average difference between the reference trajectory per image, (2) Translational Relative Pose Error (RPEtrans), and (3) Rotational Relative Pose Error (RPErot) which measure the deviation from the reference and estimated trajectories in terms of translational and rotational motions, respectively. ATE is expressed as follows:

$$ATE = d(x,y) \quad (1)$$

$$d(x,y) = ((x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2)^{\{\frac{1}{2}\}} \quad (2)$$

where d(x,y) is the Euclidean distance, x and y represent the estimated global position of the camera and its reference counterparts, respectively. Since both the reference and estimated plots are independent, a solid transformation in x is used to form a solution that maps the estimated trajectories to y reference trajectories. RPEtrans and RPErot are described as follows:

$$A = (Y_i^{-1}.Y_{i+1})^{-1} \cdot (X_i^{-1}.\ X_{i+1}) \quad (3)$$

$$Trans.RPE = (A_{0,3}^2 + A_{1,3}^2 + A_{2,3}^2)^{\frac{1}{2}} \quad (4)$$

$$x = \frac{1}{2}(A_{0,0} + A_{1,1} + A_{2,2} - 1) \quad (5)$$

$$Rot.RPE = \arccos(max(min(x,1),-1)) \quad (6)$$

where A is the relative pose error and X and Y are the estimated and reference poses, respectively. In Figure 3, the evaluation results of the pose estimation for both scenarios are quantitatively presented with the mean and standard deviation of these metrics. The results indicate that the virtually pre-trained models perform better with 22.18 and 34.34 [cm] ATE for 120.14 and 116.99 [cm] trajectories in small intestine and colon compared to the models trained only on real data with 33.28 and 53.42 [cm] ATE for the same trajectories, respectively. Accordingly, it can be observed that pre-training the neural network model with synthetic dataset enhances the performance of the algorithm in the final pose estimation.

Organ	Small Intestine		Colon	
Trajectory Length [m]	1.2028		1.1699	
Training Type	Pre-training with Synthetic Data	Without Pre-training With Synthetic Data	Pre-training with Synthetic Data	Without Pre-training With Synthetic Data
ATE (mean±std) [m]	0.2218±0.2155	0.3328±0.1845	0.3434±0.1757	0.5342±0.4408
Trans. RPE (mean±std) [m]	0.0019±0.0037	0.0026±0.0024	0.0022±0.0026	0.0048±0.0042
Rot. RPE (mean±std) [°]	1.3249±1.1247	3.1951±3.892	1.3258±0.9865	4.3228±4.6234

Figure 3. Pose estimation results in real endoscopy data. The Endo-SfMLearner neural network model is trained in two different scenarios. In the first case, only real endoscopy images from the EndoSLAM dataset are used for training. In the second case, the model is fine-tuned with real data after pre-training with synthetic data. Both models are tested with small intestine and colon trajectories of the real dataset and ATE, translational and rotational RPE scores are used for quantitative evaluations. The numerical results show that virtual pre-training improves the approximation of the final poses in all test cases

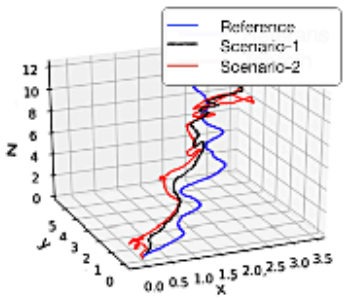
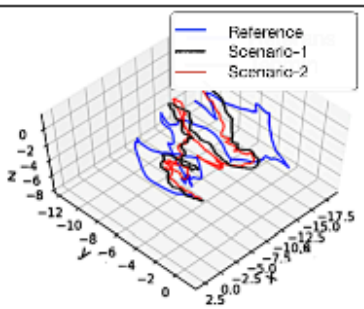
	Stomach			
	Synthetic Data		EndoSLAM	
Estimations				
Trajectory Length [m]	0.1651		0.7323	
Training Type	Scenario-1	Scenario-2	Scenario-1	Scenario-2
ATE (mean±std) [m]	0.3017±0.1713	0.2421±0.1926	0.3522±0.1233	0.2876±0.0775
Trans. RPE (mean±std) [m]	0.0065±0.0001	0.0042±0.0001	0.0015±0.0009	0.0014±0.012
Rot. RPE (mean±std) [°]	3.7322±3.5797	3.4230±3.3252	4.2906±3.7814	3.9400±3.6539

Figure 4. Pose estimation results in synthetic and real endoscopy data. The Endo-SfMLearner model pre-trained with synthetic data was tested in the stomach trajectory of both synthetic and EndoSLAM datasets. In the first scenario, 300 synthetic and 1450 real endoscopy images are used for training while the second scenario consists of 600 synthetic and 900 real endoscopy images. The results are shown qualitatively with the reference and estimated trajectory plots whereas quantitatively with ATE, translational and rotational RPE. As can be seen, the number of synthetic data used in the training set improves the performance in pose estimation for both datasets.

As a further investigation, the Endo-SfMLearner model is trained in two different scenarios to test the effect of the number of synthetic data used in the pre-training set. In the first case, the model is pre-trained with 300 synthetic images and fine-tuned with 1450 real endoscopy images from the EndoSLAM dataset. In the second scenario, the number of synthetic images in the pre-training set is increased to 600 while the number of real endoscopy images in the fine-tune set is reduced to 900. Each model is tested on both synthetic and real endoscopy images of the stomach, and the results are presented qualitatively and quantitatively in Figure 4. The predicted trajectory curves in scenario two are more able to follow the ground truth trajectories in both real and synthetic test data, which is confirmed by the quantitative results showing the efficiency of synthetic data based training for situations where less real data is available and more synthetic data is used for training the neural network.

Discussion

In this study, fully-labelled and realistic synthetic endoscopic images were generated in a virtual environment corresponding to the topology and tissue of the organs of the gastrointestinal tract to facilitate the application of deep learning based algorithms for simulation to real transfer. The performance of a state-of-the-art method was tested on two different tasks using synthetic data in the training phase. Both qualitative and quantitative results showed that the models trained on synthetic data can perform successfully on real medical data, even when the training set consists only of synthetic images. Accordingly, the use of synthetic data improves the performance of deep learning models, which can increase the efficiency of studies related to endoscopy by addressing common technical problems in surgical operations. As a future study, we plan to create disease classes in the virtual environment to analyse the quality of the generated images in terms of anatomical features and diversity, and to incorporate modalities for segmentation and classification problems.

Conflict of interests
The authors declare that they have no competing interests.

Financial Disclosure
This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with grant 2232 - The International Fellowship for Outstanding Researchers

Ethical approval
Ethics Committee Approval: Ethics committee approval was not obtained for this study. As no human and experimental studies were conducted in this study.

References

1. Zhengrong L, Robert R. Virtual colonoscopy vs optical colonoscopy. Expert Opin Med Diagn. 2010;4:159–69.
2. Potter J, Slattery M. Colon Cancer: A review of epidemiology. Epidemiologic Reviews. 1993;15:499–45.
3. Luo XB. ,Hui-Qing Zeng,Yan D, Xiao C, et al. A novel endoscopic navigation system: simultaneous endoscope and radial ultrasound probe tracking without external trackers.In: MICCAI. 2019;47-55.
4. J. Peng, X J. Peng, H. Tang, et al. Design and fabrication of an integrated convex ultrasound endoscope for digestive tract imaging. IEEE International Ultrasonics Symposium (IUS). 2015;1-4.
5. Li J, Liu DX, Liu ZB , Wei J. Imageological study on esophageal carcinoma at clinical stages and significance of tumor markers MMP-9 and NGAL. Biomedical Research, Special Issue. 2017;568-70.
6. Merve T¸ Bülent Y. Super resolution convolutional neural network based pre-processing for automatic polyp detection in colonoscopy images. Computers and Electrical Engineering. 2021;90:1-11.
7. Takemoto T, Yanai H, Tada M, et al. Application of Ultrasonic Probes Prior to Endoscopic Resection of Early Gastric Cancer. Endoscopy. 1992;24:329-33.
8. Davies RJ , Richard M, Nicholas C. Colorectal cancer screening: prospects for molecular stool analysis. Nature Reviews Cancer. 2005;5:199-09.
9. Meining A. , Semmler A, Kassem M, et. al. The effect of sedation on the quality of upper gastrointestinal endoscopy: an investigator-blinded, randomized study comparing propofol with midazolam. In:Endoscopy. 2007;39:345-49.
10. Sehyuk Y, Metin S. Design and rolling lo-comotion of a magnetically actuated soft capsule en-doscope. IEEE Transactions on Robotics. 2011;28:183-94.
11. Li Y, Sixou B, Peyrin F. Review of the deep learning methods for medical images super resolution problem. IRBM. 2020;42:120–33.
12. Lee J, Young J, Yoon W, et al. Spotting malignancies from gastric endoscopic images using deep learning. Surg. Endosc. 2019;33:3790–7.
13. Xintao W, Ke Y, Shixiang W, et al. ESRGAN: Enhanced super-resolution generative adversarial networks. ECCV. 2019;11133:63–79.
14. Faisal M, Nicholas J Durr. Deep learning-based depth estimation from a synthetic en-doscopy image training set. Medical Imaging. 2018;10574.
15. Alyafi, B., Diaz, O., and Marti, R. DCGANs for Realistic Breast Mass Augmentation in X-ray Mammography. In: Medical Imaging 2020: computer-aided diagnosis. International Society for Optics and Photonics. 2020;1909.02062.
16. Alyafi, Basel, Diaz, Oliver, Vilanova, Joan C., del Riego, Javier and Martí, Robert. Quality analysis of DCGAN-generated mammography lesions. In: Arxiv.1911.12850v1. 2019.
17. Kağan İ, Ibrahim o, Abdulhamid O, et. al. VR-Caps: A virtual environment for capsule endoscopy. Medical Image Analysis. 2021;70.
18. Arthur J,Vincent-Pierre B, Ervin T, et. al. Unity: A general platform for intelligent agents. arXiv preprintarXiv:1809. 02627, 2018.
19. François F, ,Christian D, Hervé D, et. al. SOFA: A Multi-Model Framework for interactive physical simulation. softtissue biomechanical modeling for computer assisted surgery. ed. by yohan payan, vol. 1, studiesin mechanobiology, tissue engineering and bioma-terials. Springer 2012;283–21.
20. FR[dataset] Lucchesi and Aredes ND. RadiologyData from The Cancer Genome Atlas Stomach Ade-nocarcinoma [TCGA-STAD] collection. Cancer Imaging Arch. 2016;10, K9.
21. Smithetal K. Data From CT COLONOGRAPHY. The Cancer Imaging Archive. 2015.
22. Konstantin P, Kristin R, Carsten G, et. al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. Proceedings of the 8th ACM on Multimedia Systems Conference. 2017;164–9.
23. Kutsev B, Guliz Irem G, Gulfize C, et. al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: endo-SfMLearner. Medical Image Analysis. 2021;71.