EndoVAE: Generating Endoscopic Images with a Variational Autoencoder

Conference Paper · June 2022

DOI: 10.1109/IVMSP54334.2022.9816329

3 authors, including:

Dimitris Diamantis
Technological Educational Institute of Lamia
25 PUBLICATIONS **413** CITATIONS

SEE PROFILE

Dimitris Iakovidis
University of Thessaly
235 PUBLICATIONS **5,490** CITATIONS

SEE PROFILE

EndoVAE: Generating Endoscopic Images with a Variational Autoencoder

Dimitrios E. Diamantis, Panagiota Gatoula, Dimitris K. Iakovidis
Department of Computer Science and Biomedical Informatics
University of Thessaly, Lamia, Greece
{didiamantis, pgatoula, diakovidis}@uth.gr

***Abstract*—The generalization performance of deep learning models is closely associated with the number and diversity of data available upon training. While in many applications there is a large number of data available in public, in domains such as medical image analysis, the data availability is limited. This can be largely attributed to data privacy legislations, including the General Data Protection Regulation (GDPR), and the cost of data annotation by experts. Aiming to address this issue, data augmentation approaches employing deep generative models have emerged. Existing augmentation techniques are primarily based on Generative Adversarial Networks (GANs). However, ill-posed training issues of GANs such as nonconvergence, mode collapse and instability in conjunction with their demand for large scale training datasets, complicate their use in medical imaging modalities. Motivated by these issues, this paper investigates the performance of alternative generative models *i.e.*, Variational Autoencoders (VAEs) in endoscopic image synthesis tasks. Contrary to the conventional GAN-based approaches that aiming at augmenting the existing endoscopic datasets the proposed methodology constitutes feasible the complete substitution of medical imaging datasets from real individuals with artificially generated ones. The experimental results obtained validate the effectiveness of the proposed methodology over the state-of-art.**

Keywords—Wireless Capsule Endoscopy, Variational Autoencoders, Medical Image Synthesis

I. INTRODUCTION

Over the last decade an alarming increase of diseases associated with the Gastrointestinal (GI) tract, is observed. The preliminary detection and the accurate diagnosis of pathological findings is considered vital for the prevention GI tract diseases [1]. Wireless Capsule Endoscopy (WCE) is regarded one of the most prominent and less invasive screening techniques for the examination of the GI tract and the detection of possible anomalies, such as polyps, vascular and inflammatory lesions [2]. WCE employs a swallowable capsule with an embedding camera that traverses through the GI tract and produces a video that approximately consists of more than 60.000 frames. An enormous amount of WCE examinations is evaluated from the physicians for the detection and identification of possible pathologies. Understandably, the above procedure is a challenging task and certainly prone to human errors even for experienced physicians [2]. To overcome those difficulties computer aided (CAD) systems have been employed to assist the diagnosis and /or treatment process of WCE examinations [3].

Over the past few years, recent advances in deep learning have entrained the field of medical imaging [4]. In this context, the development of intelligent CAD systems oriented to endoscopic image analysis consists a subject of extensive research [5–7]. Data availability is considered an integral component for development of an automated endoscopic image analysis system. However, data privacy legislations, such as the General Data Protection Regularization (GDPR) [8] introduce restrictions and delays in the acquisition and sharing of medical images, even for scientific and research purposes. Furthermore, the existing publicly available endoscopy databases are limited, usually imbalanced with the regards to the pathological findings and they are insufficient in terms of diversity [4].

To mitigate these issues, augmentation techniques primarily exploiting deep learning methodologies have been proposed recently. The emerging methodologies mainly rely on deep generative models, such as Generative Adversarial Networks (GANs) [9]. However, the fact that GANs suffer from various training problems, such as inability to converge, training instability, and mode collapse, in conjunction with the fact that they conventionally require an adequate amount of diverse training samples, operates as a limiting factor concerning their performance [10]. The results of current methodologies for endoscopic image synthesis [11–13] indicate that there are still a lot of open issues, mainly with respect to the quality and the diversity of the synthetic images, affecting the degree to which they can be considered as realistic.

This study aims to address the limitations of the conventional generative approaches by investigating the use of a Variational Autoencoder [14], for plausible endoscopic image synthesis. Therefore, the contributions of our work are summarized as follows:

- The novel introduction of a VAE architecture for reasonable endoscopic image synthesis with sufficient quality and diversity aiming at the total substitution of the real endoscopic images and not for the common augmenting purposes.
- A performance assessment study of the proposed methodology on a publicly available endoscopic database [15], validating its effectiveness for GI abnormality detection in endoscopy images, in comparison to a relevant state-of-the-art approach.

We acknowledge support of this work by the project “Smart Tourist” (MIS 5047243) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund)

The rest of the paper is organized in four sections as follows: Section II presents a synopsis with respect to the deep generative models for endoscopic image synthesis. Section III describes the proposed methodology, and Section IV presents the experimental setup and the results obtained. The last section summarizes the conclusions derived from this study.

II. Related work

Deep generative models, such as GANs [9] consist a conventional powerful tool regarding high-quality synthesis of natural images. However, their applicability in the context of medical imaging modalities, such as endoscopy images has been proven to be a demanding task. This can be attributed to the fact that endoscopy images, contrary to other medical modalities do not follow a formal protocol and lack certain structure as well.

Existing approaches for endoscopic image synthesis are primarily built upon GANs. In [13], a GAN framework conditioned on a combined input of an edge-filtering and a mask for synthesizing polyp images to enhance the polyp detection in colonoscopy videos, was proposed. Following this work, the majority of the studies conducted synthesis focused on the generation of colonoscopy images with polyp lesions due to the lower complexity and data availability of the synthesis task concerning the particular modality. The work presented in [12] proposed a patch-based approach for embedding polyp findings into normal endoscopy images. In [11] an existing GAN model [16] repurposed for augmentation of endoscopic datasets was proposed. The aim of that model was to synthesize false negative colonoscopy images and further enrich them with an adversarial attack process, to re-train different anomaly detection models. Recently, in [17] an adapted GAN architecture, originally proposed for portrait image generation, was used to synthesize endoscopic images. The GAN was trained on a large-scale private dataset and provided satisfactory results in terms of quality for improving polyp lesion detection, including lesions that were barely untraceable by the human eye. The work of [5] was based on a hybrid GAN-VAE framework [18] to handle the imbalance of endoscopic datasets. The proposed hybrid method operates as an augmentation technique to improve the real-time classification of small bowel findings.

In the context of generating WCE endoscopic images of the small bowel with and without inflammatory conditions, the work of [19] used a non-stationary texture synthesis GAN. Moreover, that work was the first one suggesting that fully synthetic datasets can be leveraged to solely train state-of-the-art CNN endoscopic classifiers with sufficient generalization abilities.

III. Methodology

The methodology proposed in this paper, presents a novel VAE architecture, namely Endoscopic VAE (EndoVAE), for generating plausible endoscopic images with and without inflammatory conditions. A VAE model [14] is composed of two parts: an encoder and a decoder. The encoder maps an input volume x to a compressed latent representation z, and the decoder reconstructs the input volume based on the latent representation of the encoder. The encoder of a VAE, contrary to the conventional autoencoders that compress the input volume to a latent vector, maps the input volume to two distinct vectors that correspond to the mean μ, and standard deviation σ, parameters of a Gaussian distribution $\mathcal{N}$. The VAE model is trained by minimizing the reconstruction error between the input volume x and the reconstructed representation of the decoder $\hat{x}$. Additionally, to ensure that the encoder compresses the input volume into a latent representation of a Gaussian Distribution, the reconstruction loss is combined with the Kullback-Leibler (KL) Divergence which operates as a regularization term. It has been proven [14] that the total loss of a VAE model can be formulated as follows :

$$\mathcal{L}_{VAE} = D_{KL}\big(\mathcal{N}(\mu,\sigma)||\,\mathcal{N}(0,\mathrm{I})\big) + \frac{1}{L}\sum_{l=1}^{L}\log p(x|z_l\,;\vartheta) \quad (1)$$

where the first term corresponds to the KL – divergence, whereas the second term represents the reconstruction error. L refers to the sample size of the Monte Carlo method sampling from the approximate distribution of the encoder, $p(x|z_l\,;\vartheta)$ describes the conditional probability distribution over x given z and ϑ denotes the parameters of the decoder.

The architecture of the proposed VAE synthesis model is illustrated in Fig. 1. The encoder of our model consists of six consecutive convolutional layers. The first two layers consist of 16 and 32 filters respectively. The third layer also contains 32 filters, and this number is doubled for every next layer.

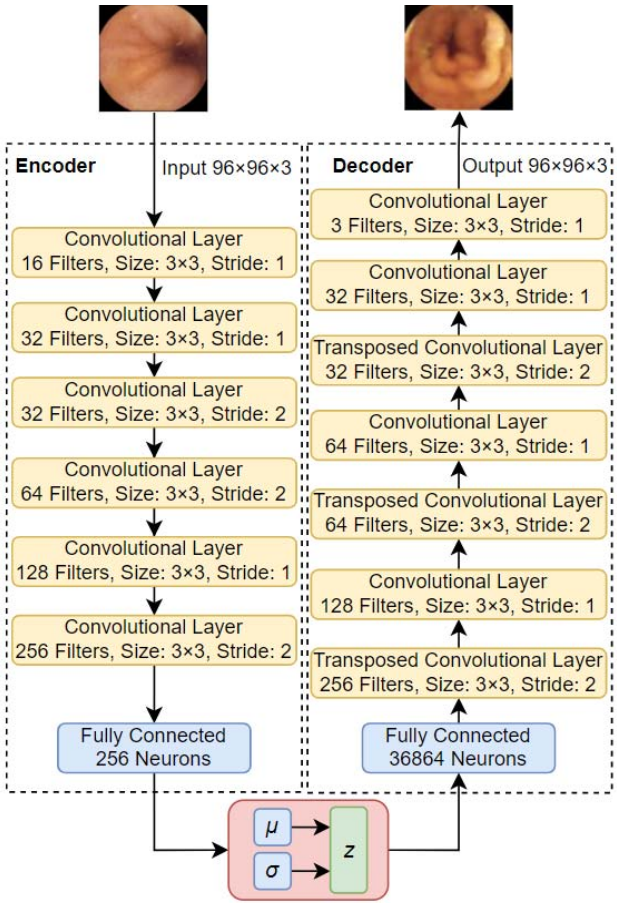

Fig. 1. Visual reprepesentation of the proposed EndoVAE methodology.

The output volume of the convolutional layers is flattened, and it passes through a fully connected layer composed of 256 neurons. At the end of the encoder arranged two distinct fully connected layers each one separately attached to the layer with the 256 neurons. These fully connected layers consist of six neurons with for estimating the distribution parameters μ, σ of the latent space.

The decoder part of the proposed model consists of a fully connected layer that contains 36,864 neurons followed by seven transpose convolutional layers with 256, 128, 64, 64, 32, 32 and 3 filters respectively. All the convolutional layers in the encoder and the decoder part of our model perform 3 × 3 convolutional operations. Rectified Linear Units (ReLUs) have been selected as activation functions for all the layers of the encoder and the decoder, except from the last transpose convolutional layer that uses the sigmoidal activation function. These hyper-parameters were empirically selected.

IV. Experiments and results

A. Dataset and training

The EndoVAE methodology was evaluated on the publicly available WCE database KID [15]. KID database consists of 2371 WCE high quality anonymized RGB images of 360×360 pixels size, from the entire GI tract. More specifically, it includes 1778 normal images from the esophagus, the small bowel and the colon along with 574 images of abnormal findings. Abnormal images represent a variety of pathological conditions related to the small bowel such as polypoid and ampulla lesions apart from vascular and inflammatory findings. We performed two-phase experiments on a normal and an abnormal subset of the KID dataset, to evaluate the effectiveness of the proposed VAE architecture in the respective image synthesis tasks. More specifically, the generation of endoscopic images of the small bowel was assessed using a set of 728 images of normal tissue (normal subset), and a set of 227 abnormal images of inflammatory conditions (abnormal subset). We applied the same preprocessing procedure for both training subsets. Particularly, the RGB images were cropped, resized to a spatial resolution of 96×96 pixels, and normalized to the interval of [0,1]. No other data augmentation technique was applied to the training data. The Adam optimizer [20] with an initial learning rate of 1×10-3 was selected to update the parameters of the proposed model. The model was trained for 5000 epochs in total, and the batch size was set to 128.

B. Evaluation methodology

To validate the performance of the proposed VAE-based framework in endoscopic image generation, we investigate the performance of a state-of-the-art endoscopic classifier trained solely on the images artificially generated by the proposed methodology. In particular, the LB-FCN *light* [21] model was employed to conduct two separate experiments. First, it was trained on the real normal and abnormal subsets of KID dataset and then its performance was compared following the same training procedure with images exclusively synthesized by the VAE architecture presented. In both experiments the LB-FCN *light* model was evaluated on the same real images. The number of synthetic images used for training was equal to that of the real subsets in the first experiment.

LB-FCN *light* is a lightweight version of the LB-FCN architecture [6] which is a Fully Convolutional Neural Network [22], originally proposed for abnormality detection in WCE images of the GI tract. The principal components of this network are the multi-scale feature extraction modules, connected with large residual connections, that decisively contribute to the generalization ability of the model even towards to small-scale endoscopic datasets. Furthermore, this model is characterized by a relatively low number of free parameters in comparison with relevant CNN architectures. This fact is mainly attributed to the separable depth-wise convolutions that leverages.

To attenuate a potential selection bias during the evaluation procedure we adopt a stratified 10-fold-cross-validation training process. Therefore, real and synthetic data were split into 10 disjoint subsets, where 9 of them was used for training the LB-FCN *light* architecture and 1 of them was used to assess the trained model. It should be noted that, the same test set used for the evaluation of the classifier trained with real images, was also used to assess the classification performance when trained with the synthetic images. This process was repeated 10 times and each time a different subset is kept for testing. For training the LB-FCN *light* we follow the training optimization settings of [19].

To evaluate the classification performance of the trained LB-FCN *light* models, Receiver Operating Characteristic (ROC) curves were calculated. ROC curves are considered as indicators of the diagnostic ability of classification models. They display a tradeoff between True Positive (TPR) and False Positive (FPR) Rates for a variety of decision thresholds. The Area Under ROC (AUC) measure [23] was selected to quantify the classification performance of the models trained on synthetic and real datasets. In this case AUC is regarded as a more representative metric for estimating the classification performance considering that the real endoscopic images are unevenly distributed between the two classes *i.e.*, normal and abnormal.

C. Comparative study and results

The experiments showed that training the LB-FCN *light*, exclusively on synthetic images generated by EndoVAE, results into an 81.9±0.9% AUC. When the same process was performed on the real images of KID dataset AUC value reached 90.9±0.8%. These results show that the proposed methodology effectively enhances the classification performance of the WCE images in comparison to the previous GAN-based approach [19] that resulted into 79.1±0.7% AUC. Furthermore, the results obtained by EndoVAE are comparable with other methodologies evaluated using real images on the same dataset. For instance, the BoW-based methodology of [24], which is based on features extracted from the CIE-Lab color space, resulted in a classification performance ranging from 77% to 81%.

Figs. 2 and 3, present a qualitative comparison of the images synthesized by the proposed VAE architecture and the original images of the KID dataset. From Fig. 2 it can be observed that EndoVAE effectively synthesizes images that naturally depict distinctive features of endoscopy images. Additionally, it can be

Fig. 2. Sample images from the KID WCE Database. The first row consists of healthy small bowel images and the second images with various inflammatory conditions.

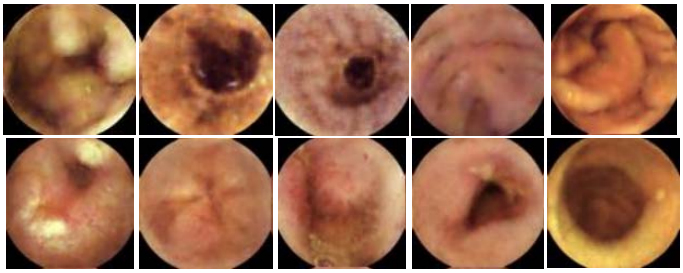

Fig. 3. Sample images generated using the proposed VAE architecture. The first row consists of healthy small bowel images and the second images with various inflammatory conditions.

observed that artificially generated images are characterized from sufficient diversity. Moreover, synthesized images seamlessly embed inflammation conditions in the endoscopic background. However, the generated images suffer from blurriness and in some cases, colors observed subdued compared to the real images. Comparing the results of EndoVAE with the results of the state-of-the-art GAN-based methodology for endoscopic synthesis of inflammatory conditions [19], it can be observed that the proposed synthesis results in more stable training procedure, crispier and more diverse images.

V. Conclusion

The originality of the proposed methodology namely Endoscopic VAE (EndoVAE), lies in the fact that it investigates the effectiveness of an alternative deep generative model, *i.e.*, VAEs, with respect to endoscopic image synthesis. To the best of our knowledge this is the first study in the context of endoscopic image generation that employs a such an alternative model instead of a GAN. Furthermore, unlike most of the previous relevant studies, and towards the vision set in our previous study [19], image generation is not considered simply in the context of a conventional data augmentation task, but in the context of the complete substitution of real images with synthetic ones.

The promising results obtained from the experimental evaluation of the proposed methodology certainly demonstrate that VAEs can efficiently generate realistic endoscopic images. The outcomes of our investigation indicate that VAE-based methodologies can improve in a substantial way conventional approaches that are exclusively build upon GANs and their variants. Although, the classification performance when trained with real images yields better results, the findings of this study confirm that solely artificially generated medical imaging datasets could substitute those comprised of data acquired from real patients regarding the development of intelligent CAD systems with sufficient generalization abilities. An aspect that has space for improvement concerns the quality of synthesis. This is undoubtedly a limitation of most of the existing generative models for endoscopic image synthesis. In our future research we indent to concentrate on the improvement of the quality of synthesis on the basis of the promising findings presented in this paper. To this direction we are planning to improve the architecture of the proposed EndoVAE model for effectively dealing with the blurriness, color consistency and the resolution of synthetic images and conduct experiments in publicly available datasets of various sizes.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, 2021, pp. 209–249.

[2] A. Riphaus, S. Richter, M. Vonderach, and T. Wehrmann, "Capsule endoscopy interpretation by an endoscopy nurse - a comparative trial," *Z Gastroenterol*, vol. 47, Mar. 2009, pp. 273–276.

[3] M. Liedlgruber and A. Uhl, "Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review," *IEEE reviews in biomedical engineering*, vol. 4, 2011, pp. 73–88.

[4] R. Pannala, K. Krishnan, J. Melson, M.A. Parsi, A.R. Schulman, S. Sullivan, G. Trikudanathan, A.J. Trindade, R.R. Watson, J.T. Maple, and D.R. Lichtenstein, "Artificial intelligence in gastrointestinal endoscopy," *VideoGIE*, vol. 5, Nov. 2020, pp. 598–613.

[5] J. Ahn, H.N. Loc, R.K. Balan, Y. Lee, and J. Ko, "Finding Small-Bowel Lesions: Challenges in Endoscopy-Image-Based Learning Systems," *Computer*, vol. 51, May. 2018, pp. 68–76.

[6] D.E. Diamantis, D.K. Iakovidis, and A. Koulaouzidis, "Look-behind fully convolutional neural network for computer-aided endoscopy," *Biomed. Signal Process. Control.*, vol. 49, 2019, pp. 192–201.

[7] G. Dimas, D. Iakovidis, and A. Koulaouzidis, "MedGaze: Gaze Estimation on WCE Images Based on a CNN Autoencoder," *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2019, pp. 363–367.

[8] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, 2017, pp. 10–5555.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets In: Advances in Neural Information Processing Systems (NIPS)," 2014.

[10] D. Saxena and J. Cao, "Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions," *ACM Comput. Surv.*, vol. 54, May. 2021.

[11] F. He, S. Chen, S. Li, L. Zhou, H. Zhang, H. Peng, and X. Huang, "Colonoscopic Image Synthesis For Polyp Detector Enhancement Via Gan And Adversarial Training," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1887–1891.

[12] T. Kanayama, Y. Kurose, K. Tanaka, K. Aida, S. Satoh, M. Kitsuregawa, and T. Harada, "Gastric Cancer Detection from Endoscopic Images Using Synthesis by GAN," 2019, pp. 530–538.

[13] Y. Shin, H.A. Qadir, and I. Balasingham, "Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance," *IEEE Access*, vol. 6, 2018, pp. 56007–56017.

[14] D.P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *CoRR*, vol. abs/1906.02691, 2019.

[15] A. Koulaouzidis, D.K. Iakovidis, D.E. Yung, E. Rondonotti, U. Kopylov, J.N. Plevris, E. Toth, A. Eliakim, G.W. Johansson, W. Marlicz, and others, "KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes," *Endoscopy international open*, vol. 5, 2017, pp. E477–E483.

[16] T.R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a Generative Model from a Single Natural Image," *CoRR*, vol. abs/1905.01164, 2019.
[17] D. Yoon, H.-J. Kong, B.S. Kim, W.S. Cho, J.C. Lee, M. Cho, M.H. Lim, S.Y. Yang, S.H. Lim, J. Lee, J.H. Song, G.E. Chung, J.M. Choi, H.Y. Kang, J.H. Bae, and S. Kim, "Colonoscopic image synthesis with generative adversarial network for enhanced detection of sessile serrated lesions using convolutional neural network," *Scientific Reports*, vol. 12, Jan. 2022, p. 261.
[18] A.B.L. Larsen, S.K. Sønderby, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *CoRR*, vol. abs/1512.09300, 2015.
[19] D.E. Diamantis, A.E. Zacharia, D.K. Iakovidis, and A. Koulaouzidis, "Towards the Substitution of Real with Artificially Generated Endoscopic Images for CNN Training," *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2019, pp. 519–524.
[20] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[21] D. Diamantis, D.-C. Koutsiou, and D. Iakovidis, "Staircase Detection Using a Lightweight Look-Behind Fully Convolutional Neural Network," 2019, pp. 522–532.
[22] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
[23] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, 2006, pp. 861–874.
[24] M. Vasilakakis, D.K. Iakovidis, E. Spyrou, and A. Koulaouzidis, "Weakly-Supervised Lesion Detection in Video Capsule Endoscopy Based on a Bag-of-Colour Features Model," *Computer-Assisted and Robotic Endoscopy*, T. Peters, G.-Z. Yang, N. Navab, K. Mori, X. Luo, T. Reichl, and J. McLeod, eds., Cham: Springer International Publishing, 2017, pp. 96–103.