IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 December 2023, accepted 13 February 2024, date of publication 16 February 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3366801

RESEARCH ARTICLE

This Intestine Does Not Exist: Multiscale Residual Variational Autoencoder for Realistic Wireless Capsule Endoscopy Image Generation

DIMITRIOS E. DIAMANTIS[1], PANAGIOTA GATOULA[1], ANASTASIOS KOULAOUZIDIS[2,3], AND DIMITRIS K. IAKOVIDIS[1], (Senior Member, IEEE)

[1]Department of Computer Science and Biomedical Informatics, University of Thessaly, 35131 Lamia, Greece
[2]Department of Surgery, SATC-C, Odense University Hospital and Svendborg Hospital, 5700 Svendborg, Denmark
[3]Department of Clinical Research, University of Southern Denmark, 5230 Odense, Denmark

Corresponding author: Dimitris K. Iakovidis (diakovidis@uth.gr)

This work was supported by the Project "Smart Tourist" 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) under Grant MIS 5047243. The publication of the article in OA mode was financially supported by HEAL-Link.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of the University of Thessaly.

ABSTRACT Medical image synthesis has emerged as a promising solution to address the limited availability of annotated medical data needed for training machine learning algorithms in the context of image-based Clinical Decision Support (CDS) systems. To this end, Generative Adversarial Networks (GANs) have been mainly applied to support the algorithm training process by generating synthetic images for data augmentation. However, in the field of Wireless Capsule Endoscopy (WCE), the limited content diversity and size of existing publicly available annotated datasets adversely affect both the training stability and synthesis performance of GANs. In this paper a novel Variational Autoencoder (VAE) architecture is proposed for WCE image synthesis, namely 'This Intestine Does not Exist' (TIDE). This is the first VAE architecture comprising multiscale feature extraction convolutional blocks and residual connections. Its advantage is that it enables the generation of high-quality and diverse datasets even with a limited number of training images. Contrary to the current approaches, which are oriented towards the augmentation of the available datasets, this study demonstrates that using TIDE, real WCE datasets can be fully substituted by artificially generated ones, without compromising classification performance of CDS. It performs a spherical experimental evaluation study that covers both quantitative and qualitative aspects, including a user evaluation study performed by WCE specialists, which validate from a medical viewpoint that both the normal and abnormal WCE images synthesized by TIDE are sufficiently realistic. The quantitative results obtained by comparative experiments validate that the proposed architecture outperforms the state-of-the-art.

INDEX TERMS Clinical decision support systems, endoscopy, gastrointestinal tract, image synthesis, variational autoencoders.

I. INTRODUCTION

Gastrointestinal (GI) tract diseases constitute a significant cause of mortality and morbidity, resulting in adverse economic effects on healthcare systems [1]. Early-stage detection and precise diagnosis of pathological conditions, such as inflammations, vascular conditions, or polypoid lesions, are critical for preventing such diseases. Among the methods facilitating the screening of the GI tract, Wireless Capsule Endoscopy (WCE) is one of the eminent options mainly due to its non-invasive nature. Contrary to conventional techniques, such as Flexible Endoscopy (FE), WCE is

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

performed using a swallowable, pill-sized capsule equipped with a miniature camera. The capsule traverses throughout the GI tract recording an RGB video, which is subsequently reviewed by specialized endoscopists. Since such a video is typically comprised with more than 60,000 frames, its assessment is demanding. The evaluation of the recorded WCE videos usually requires 45-90 minutes, and it is prone to human errors even by experienced clinicians. Aiming to mitigate the risk of such errors, various image-based Clinical Decision Support (CDS) systems have been proposed [2], [3], [4], [5], [6]. A significant aspect concerning the performance of CDS systems is their generalization ability. The availability and diversity of training data directly impact the generalization capability of such systems [2]. Although there are several publicly available annotated datasets for non-medical applications, many of which are quite large, *e.g.*, ImageNet [7], in the medical imaging domain, privacy regulations, such as the General Data Protection Regulation (GDPR) [8], make medical data acquisition challenging, even when their use is destined for research. Moreover, the amount of time and the cost of medical data annotation adversely contribute to their availability [2]. Regarding WCE, the existing open annotated datasets are still limited, they are generally smaller than other datasets, often characterized by low diversity, as they contain many similar images with a narrow range of abnormality types, and in most cases, they are highly imbalanced [9]. Consequently, the use of such datasets for training of contemporary deep learning-based CDS systems limit their effectiveness for detection and characterization of abnormalities [9].

To address these issues, conventional image augmentation techniques, enriching the datasets with rotated, translated, and scaled versions of the training images, have been employed to enhance the generalization ability of CDS systems [10]. More recently, approaches relying on deep neural networks have been investigated to increase the number of training images further by generating synthetic images. Generative Adversarial Networks (GANs) [11] are considered as a standard option for image synthesis tasks. In applications such as natural images or portrait synthesis, where data availability is not an issue, the performance of GANs is remarkable [12], [13], [14]. However, in many cases, where data availability is limited, the applicability of GANs implies various training problems, including mode collapse, unconvergence and instability [15].

Current studies [16], [17], [18], [19], indicate that still the problem of endoscopic image synthesis is far from being resolved. In most cases, synthetic images include artifacts attributed to imbalanced or insufficient training data [10], whereas the plausibility of the resulting synthetic images remains the main challenge [20]. Currently, synthetic images produced by GAN models tend not to naturally depict the structure of endoscopic tissue, and in some cases [17], [19], not to reasonably reproduce the texture and color characteristics [10]. Another issue in endoscopic image synthesis is that the appearance of lesions varies, making it difficult for the generative algorithms to learn and reproduce their characteristics using small datasets. For example, some lesions, such as inflammations are usually flat or excavated, characterized by soft or more intense color gradations, whereas others, such as polypoid lesions may be flat, sessile or pedunculated. In fact, most of the current studies [10], [16], [17], [19], [21], [22] focus on polyp image synthesis. However, the existing results indicate that most methodologies struggle with the natural integration of pathologies into the endoscopic background. Although, in current studies the artificially generated images are considered sufficient to assist data augmentation in classification and detection tasks, they often lack clinical evaluation by endoscopy experts [16], [17], [19], [21], [22] which does not validate their diagnostic/clinical value. Aiming to cope with these issues, in this study a novel approach to the generation of synthetic WCE images is proposed, based on the concept of Variational Autoencoders (VAEs) [23]. Its contributions can be summarized as follows:

- It proposes a novel VAE architecture named 'This Intestine Does not Exist' (TIDE), which combines multiscale feature extraction and residual learning, to capture feature-rich representations of the input volume, enabling training on a small number of samples. To the best of our knowledge the combination of multiscale feature extraction with residual learning has never been used in the context of VAEs for image synthesis.
- It applies TIDE in the context of WCE image synthesis aiming to fully substitute real training sets with synthetic ones. Studies from other researchers [24] have investigated WCE image synthesis only in the context of data augmentation, where just a subset of the training set was composed of synthetic images.
- It performs a spherical experimental evaluation study that covers quantitative and qualitative aspects, including a user evaluation study performed by WCE specialists, which verified that it is very hard to distinguish the synthetic datasets from the real ones.

Alongside this study, a demonstration website,[1] has been created aiming to present the performance of the TIDE openly, and to become the first publicly available real-time intestine dataset generation platform. It is worth noting that the platform was developed using the Algorithm-agnostic architecture for Scalable Machine Learning (ASML), that we proposed in [25].

The rest of this paper is organized into four sections. Section II outlines the contribution of generative models in the medical imaging domain emphasizing the synthesis of endoscopic images. Section III presents the proposed VAE architecture for image generation. Section IV describes the evaluation methodology and includes comparative results obtained from the conducted experiments. Insights of this

[1] https://this-intestine-does-not-exist.com

study are discussed in Section V, and in the last section, conclusions are drawn, and future directions are suggested.

II. RELATED WORK

In medical imaging, synthetic image generation has stimulated great scientific interest and several studies have been conducted in a variety of contexts, mainly using GANs. Most of the renowned GAN architectures have been applied for medical image synthesis. For instance, deep convolutional GANs have been used to generate plausible brain MRI images [26]. In the spirit of conditional image generation, Mahapatra et al. [27] expanded the Pix2Pix framework [14] for the production of realistic-looking chest X-ray images with nodules, based on manually segmented regions. Jin et al. [28] expanded that framework for synthesizing 3D nodules in CT images. Another widely used adversarial framework, called Progressive Growing GAN (PGGAN) [29], was trained for the generation of convincing dermoscopic images with skin lesions [30]. A multiscale GAN was proposed in [31], aiming to the fusion of different MRI modalities into a single synthetic image with richer diagnostic information for the clinicians. That study showed that multiscale information can significantly enhance the quality of the generated images. The adversarial learning scheme proposed in [32], called CycleGAN, has been used for cross-modality medical image synthesis in [33], where it was applied for unpaired image-to-image translation between MRI and CT modalities of brain images. Cai et al. [34] modified the standard CycleGAN framework for supporting simultaneous 3D synthesis and segmentation between MRI and CT modalities of cardiac and pancreatic images, while preserving the anatomical structures. In [35], the Unsupervised Image-to-image Translation (UNIT) VAE-GAN framework [36] was applied to generate eye fundus images from segmented vessel trees. Hirte et al. [37] applied another variation of a hybrid VAE-GAN model [38], to generate realistic MR brain images with improved diversity. Another architecture, called Residual Inception Encoder-Decoder Network (RIED-Net) was inspired by U-Nets [39]. In that work the residual connections improved the fusion of images from two different modalities, aiming to assist breast cancer and Alzheimer's disease detection.

In the field of endoscopy, generating realistic images has proved to be a more challenging task [10], [20], [40]. This can be explained by the fact that no specific patterns are inherent in endoscopic images [10], nor can their content be described by well-defined structures, as in the case of CT, MRI or other medical imaging modalities. The work of [19] presented a GAN conditioned on edge-filtering combined with a mask input to synthesize images. That study was focused on polyp image generation, aiming to improve polyp detection in colonoscopy videos. However, it reported limitations that include deterministic polyp generation, and insufficient variation of generated polyp features in terms of color and texture. In [17], a patch-based methodology was adopted to incorporate gastric cancer findings in normal gastroscopy images. However, as commented in [16], the positioning of the polyps in that methodology is performed manually otherwise the result can be unnatural, especially with respect to the polyp features, such as color and texture. He et al. [16] introduced a data augmentation technique based on the GAN model of [41] by following an adversarial attack process. Lately, in the work presented in [22], the generative framework proposed in [42] was adapted to produce random polyp masks, which were then combined with normal colonoscopy images to construct a conditional input. The formulated conditional input, was leveraged for training a CycleGAN model [32] to synthesize polyp images. In [43] a variation of CycleGAN was proposed to enhance of the images of a surgical simulator to resemble real intraoperative endoscopic images. In [21], a dual GAN framework conditioned on polyp masks was presented for augmenting polyp findings in colonoscopy images. However, the synthesis results in both [22] and [21] depended on the positioning of the polyp masks, which were only sometimes naturally blended with healthy endoscopic images. Recently, StyleGANv2 [13], which is a GAN architecture originally introduced for face synthesis, was used to enhance the training datasets for the detection of polyp lesions in endoscopic videos [10]. Although, that work produced realistic images in the context of polyp image synthesis, the reproducibility of its results is difficult as it relied on a private database with thousands of images available for training.

Fewer studies are dedicated to WCE image generation. In WCE, the images are of lower resolution, and the number of abnormal images is usually smaller, since the endoscopist cannot control the capsule endoscope to capture several frames of the lesions found, as in the case of FE [9]. Also, WCE is more commonly applied for the examination of the small bowel, which is very difficult to be approached by FEs, and it is invaluable for the evaluation of Inflammatory Bowel Disease (IBD), and especially of the Crohn's disease (CD) [44]. Moreover, the incidence of small bowel malignancy/neoplasia – although increasing over the last decades – remains markedly lower than colorectal or gastric neoplasia [45].

In the context of WCE image generation, Ahn et al. [24] adapted the hybrid VAE-GAN framework originally proposed in [46], to augment an existing WCE dataset, so as to improve the generalization performance of an image-based CDS system for small bowel abnormality detection. Nevertheless, the results were only indicative, not specifying the target pathological conditions, and the synthetic images suffered from blurriness, making them easily distinguishable from the real ones.

In the more complex framework of synthesizing WCE images of the small bowel, containing various inflammatory conditions, a non-stationary Texture Synthesis GAN (TS-GAN) was presented in [47]. However, the generated images had artifacts, which degraded the quality of image synthesis. These weaknesses can be partially attributed to the

limited number of training samples since the performance of GAN models typically relies on both the quantity and the diversity of the training data. To deal with this drawback, a conventional VAE, named EndoVAE, was proposed in [48] for WCE image generation. That model was composed of convolutional layers with single-scale filters in a sequential arrangement.

In this paper we propose a different VAE architecture for WCE image generation that unlike the previous ones it incorporates a multiscale feature extraction scheme and residual connections, aiming to provide WCE images of enhanced quality.

III. METHODOLOGY

The novel VAE architecture proposed in this study, named TIDE, incorporates modules, which to the best of our knowledge, have not been previously combined, in the context of variational image synthesis. The design of TIDE is based on three main principles tightly coupled with endoscopic imaging: a) *Multiscale feature extraction*, because the in a sequence of endoscopy images, the same or similar tissue structures appear at different scales, as the endoscope travels throughout the GI tract lumen; b) *Preservation of detail in image representation*, aiming to the reproduction of image features characterizing smaller lesions; c) *Model variability,* aiming to capture as much as possible of the high diversity characterizing the endoscopic image content.

TIDE uses a series of multiscale blocks (MSBs) that extract features under multiple scales, aiming to capture a feature-rich representation of the input. It is complemented by residual connections to further enhance the feature extraction scheme, and tackle the problem of vanishing gradient, encountered in deep learning models. An MSB is illustrated in Fig. 1. The purpose of these modules is twofold. Firstly, they aim to abstract image information from an input volume at various scales. Secondly, they realize a learnable fusion of the extracted information to capture more diverse representations of the input volume. Each module receives a volume of feature maps as input, which is forwarded to three parallel convolutional layers. Each of these layers has the same number of filters, and extracts features from different scales by performing convolution operations using 3×3, 5×5, and 7×7 kernel sizes with stride 1. Thus, small, medium, and large features can be extracted from the input volume. The parallel output of these layers is concatenated depth-wise, and passed forward to two additional consecutive convolutional layers, forming, in this way, a feature-rich representation of the input volume. The first of these two convolutional layers consists of a number of filters equal to the sum of the filters of the concatenated feature maps, and it performs a pointwise convolution operation, effectively facilitating the aggregation of the concatenated feature maps. into a compact representation. The purpose of the pointwise convolutional layer is to map the cross-channel correlations [49], within the concatenated feature volume which contains information from different levels of detail. The last convolutional layer of the

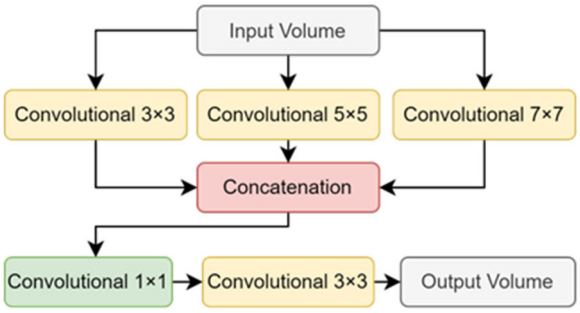


FIGURE 1. Multiscale feature extraction module.

MSB module consists of the same number of filters used in the parallel convolutional layers and performs 3×3 convolutional operations. This layer serves as a bottleneck, effectively reducing the number of feature maps, and consequently, the dimensionality along the depth axis of the input volume; thus enabling efficient feature extraction [50]. These consecutive convolutional layers perform convolutional operations with stride 1. The activation functions of all layers mentioned are the Rectified Linear Units (ReLUs).

The entire architecture of the TIDE model is illustrated in Fig. 2. It comprises two parts: an encoding network and a decoding network. The encoder receives an input volume of RGB endoscopic images with a resolution of $W\times H$ pixels, either normal or abnormal, denoted as x. The proposed encoder consists of convolutional layers destined to perform pooling operations, and to extract multiscale features. Specifically, the architecture of the encoder sequentially includes a convolutional layer with 16 filters followed by four MSBs. The first MSB consists of 32 filters, which is doubled for every next module. Convolutional layers are interposed between the consecutive MSBs to perform pooling operations by reducing the size of the intermediate feature volume to half. Those layers are composed of 64, 128, and 256, filters and perform convolutional operations with a kernel size of 3×3.

Residual connections are employed to preserve the features extracted from shallower feature extraction modules. Therefore, the input feature map volume of each module undergoes a convolutional operation with a 3×3 kernel size and a number of kernels in accordance with the number of filters leveraged by the convolutional layers of this module. The resulting feature map is aggregated with the output volume of each feature extraction module by an addition operator. Finally, the output of the residual connection is promoted to a pointwise convolution layer, preserving the number of filters of the previous convolutional layer. All the convolutional operations performed in the encoder use ReLU as an activation function.

While residual connections have been used in the past in a variety of CNN architectures, such as ResNet [51], primarily to battle the problem of vanishing gradient, they typically

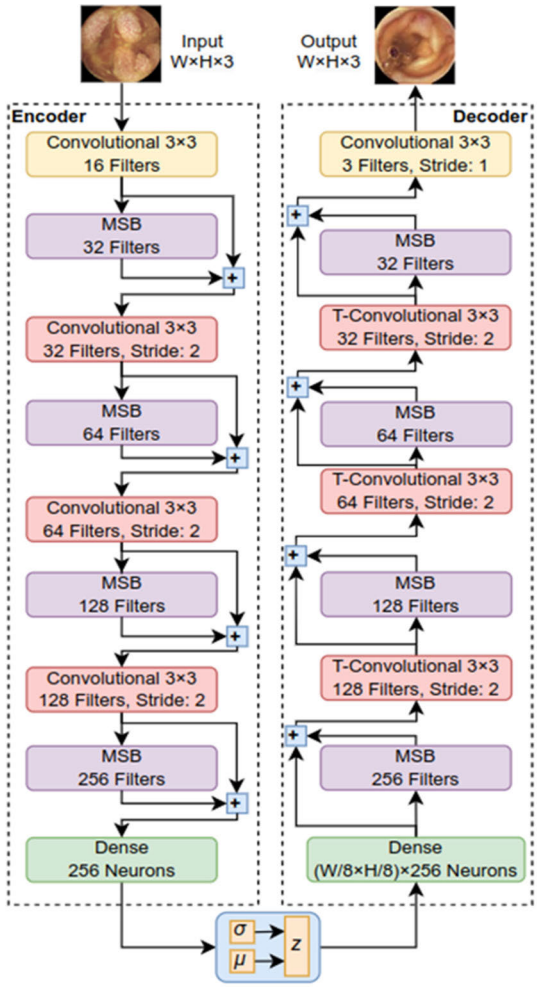


FIGURE 2. TIDE architecture.

skip just one or a few layers. In TIDE architecture, the residual connections transfer higher-level features (in case of the encoder) and lower-level features (in case of the decoder) between the MSB modules, which enable the network to battle not only the vanishing gradient problem, but also to stabilize the KL divergence. Our experiments showed that shortening the residual connections lead to rapid or exploding KL divergence while training. Regarding the MSB modules, while a similar multi-scale feature extraction technique has been used in architectures such as GoogLeNet [52], using the Inception module, they typically extract features under 2 different scales. The MSB modules used in TIDE architecture aim to capture features under 3 different scales using parallel feature extraction methodology. Parallel feature extraction procedures have also been used in CNN architectures such as ResNeXt [53], yet they extract features under a single scale. The MSB module of TIDE enables the network to capture a feature-rich representation of the input volume, which is important especially in the case of training on a small number of training samples. MSB extracts the same number of features per scale, whereas the Inception module extracts a different number of features per scale aiming to keep the free parameters of the network low in favor of computational efficiency. However, this has a negative effect on the learning capacity of the module. In the case of MSB, by keeping the same number of free parameters per scale, the learning capacity for the feature extraction process is not affected, enabling more features to be extracted, which are then efficiently reduced by the last trainable bottleneck [50] layer of MSB that completes the process. The combination of large residual connections and multi-scale feature extraction increase the generation capabilities of the proposed architecture, as it can be observed by the results of this study, when compared to conventional VAEs [36], [46], [48].

The encoder network is tasked to compress the input volume, *i.e.*, the endoscopic images, to two different latent vectors corresponding to the statistical parameters, mean μ and standard deviation σ, of a Gaussian distribution. Therefore, the output volume of the convolutional part of the encoder is flattened and directly enters a fully connected layer with 256 neurons, followed by two separated fully connected layers connected to the previous one. Each of these two layers comprises 6 neurons with no activations that estimate the parameters μ, σ of the latent space distributions.

Following this, the decoder network randomly samples a six-dimensional vector z from the distribution approximated by the encoder. Thus, the decoder, considering the latent representation z, reconstructs the input volume. At the top of the decoder's architecture, a fully connected layer resides, having 36,864 neurons. Next, the decoder adopts the architecture of the encoder, yet with an opposite order of the MSBs that, in the case of the decoder, are separated with transposed convolutional layers for performing up-sampling of the intermediate feature volume. At the end of the decoder, a transposed convolutional layer is placed, with 3 filters to predict the reconstructed RGB input volume. The spatial dimensions of the output of the decoder correspond to those of the initial volume x of endoscopic images. Consequently, the proposed VAE architecture synthesizes images of the same resolution received in the input. All the transpose convolutional operations are conducted with kernels of size 3×3 and ReLU functions as neural activations, except from the prediction layer that adopts the log-sigmoid activation function.

According to the total loss backpropagated to train a VAE model can be formulated as follows:

$$\mathcal{L}(\vartheta, \varphi; \boldsymbol{x}_i) = \mathbb{E}_{q(z|\boldsymbol{x}_i;\varphi)} \log p(\boldsymbol{x}_i|z; \vartheta) - KL\left(q(z|\boldsymbol{x}_i; \varphi)||p(z; \vartheta)\right) \quad (1)$$

where the first term corresponds to the reconstruction error of the decoder, and the second term approximates the Kullback-Leibler (KL) divergence. The KL-divergence is employed to ensure that the encoder compresses the input volume into a latent representation that follows a prior distribution $p(z; \vartheta)$. The prior distribution $p(z; \vartheta)$ is formulated as a multivariate Gaussian distribution $\mathcal{N}(z; 0, \mathbf{I})$. We let the true intractable posterior distribution $p(z|\boldsymbol{x}_i; \vartheta)$ be an approximation of the

Gaussian with an approximately diagonal covariance that is estimated according to Eq. (2):

$$\log q(z \mid \boldsymbol{x}_i; \varphi) = \log N(z; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2 \mathbf{I}) \tag{2}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i^2$ are the outputs of the encoder part of VAE.

Thus, Eq. (1) is formulated as follows:

$$\mathcal{L}(\vartheta, \varphi; \boldsymbol{x}_i) \simeq \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log \sigma_{i,j}^2 - \sigma_{i,j}^2 - \mu_{i,j}^2\right) + \frac{1}{L} \sum_{L=1}^{L} \log p(\boldsymbol{x}_i | \boldsymbol{z}_l; \vartheta) \tag{3}$$

where

$$z_l = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_l \text{ and } \boldsymbol{\epsilon}_l \sim \mathcal{N}(0, \mathbf{I}) \tag{4}$$

J denotes the dimensionality of the underlying manifold, L refers to the sample size of the Monte Carlo method sampling from the approximate posterior distribution of the encoder, φ represents the parameters of the encoder network, ϑ represents the parameters of the decoder network, and, and symbol $\odot$ represents the Hadamard product operation.

IV. EXPERIMENTS AND RESULTS

A. DATASETS AND TRAINING OF THE GENERATIVE MODEL

Considering the clinical utility of WCE and its importance for evaluating inflammatory conditions of the small bowel, two datasets, namely KID Dataset 2, and the Kvasir-Capsule dataset, were used for experimentation [54], [55]. To the best of our knowledge these datasets are the only publicly available annotated WCE datasets that include inflammatory lesions, such as erythemas, erosions, and ulcers. Both datasets are anonymized, not containing any information that would enable either patient identification or distinction between different patients. KID [54] is a WCE database designed to assess CDS systems. It includes 728 normal images and 227 images with inflammatory lesions of the small bowel, with a resolution of 360 × 360 pixels. The images were acquired using Mirocam® (IntroMedic Co., Seoul, Korea) capsule endoscopes. The Kvasir-Capsule [55] is a video capsule endoscopy dataset which includes 74 unlabeled videos and annotated image frames extracted from 43 labeled videos. The annotated images illustrate normal findings, anatomical landmarks, and various pathological findings for which bounding box masks are provided. It includes a total of 34,338 normal images and 1,519 images of inflammatory lesions, with a resolution of 336 × 336 pixels. The images were acquired using an Olympus EC-S10 endocapsule. Despite the relatively larger size of this dataset, it contains many images that are similar to each other. To reduce frame redundancy, while maintaining the two datasets equivalent in size, a subset of 728 normal and 227 abnormal representative, non-overlapping WCE images from the whole dataset, was sampled using the image mining methodology described in [56] (the filenames of the sampled images are provided as supplementary material). Different TIDE models were trained separately, on normal, and abnormal subsets of the KID and Kvasir-Capsule datasets, respectively, *i.e.*, a TIDE model was trained to generate normal images, and another one was trained to generate abnormal images, per dataset. The two datasets were not considered jointly because they were acquired using different types of capsule endoscopes. The implication of using such a joint dataset for training TIDE was experimentally investigated, and the results are presented in the next subsection. Regarding the training process, no other data augmentation techniques were applied on the training sets, as it was affecting the appearance of physiological structures. TIDE was trained using early stopping, with a maximum limit of 5,000 epochs, using batches of 128 samples. Considering its effectiveness in relevant applications, the Adam optimizer [57] was selected to train the model, using a learning rate initially set to 0.001.

For the training of the networks, we used a GPU NVIDIA RTX 3090 with 24GB GDDR6X RAM, 10,496 CUDA Cores and base clock speed at 1.8GHz. Inferences were performed on the same GPU, with an average rate of 15 milliseconds per image. Inferences are time-efficient also on CPUs. For example, the demonstration website of TIDE[1], executes inferences with an average rate of 67 milliseconds per image, on a single, not dedicated CPU core, of an Intel Core i5 processor with 4 cores, 3.8GHz and 8GB of RAM.

B. QUANTITATIVE EVALUATION

The main goal of this experimental study is to investigate if a WCE dataset composed solely of synthetic images can be effectively used to train a classifier, so that it accurately learns to discriminate real abnormal from real normal images. Therefore, the classification performance can be considered as a representative index for quantitative evaluation of the synthetic WCE datasets generated using TIDE [58]. The classification performance was quantified by examining the Receiver Operating Characteristic (ROC) curves, considering that the WCE datasets are imbalanced, and that the classification problem is binary. ROC curves indicate the diagnostic ability of a classification system by illustrating a trade-off between True Positive (TPR) and False Positive (FPR) Rates using various decision thresholds. The Area Under ROC (AUC) measure [59], was computed to assess the performance of the trained classification models, because, unlike other measures, such as accuracy, sensitivity and specificity, which are obtained using only a single decision threshold, it is insensitive to imbalanced class distributions [59], [60].

Considering that the extensive experimental work required for this study is computationally demanding, we selected LB-FCN *light* classifier [61], as a computationally more efficient version of LB-FCN, which is a state-of-the-art classifier proposed for improved classification of endoscopic images [62] that has been previously used in relevant studies [47], [48].

The experimental procedure can be outlined as follows: 1) a reference performance of LB-FCN *light* per real dataset was estimated for the classification of each dataset into

normal and abnormal (inflammation) classes; 2) the trained TIDE models were used to randomly generate different sets of synthetic normal and abnormal WCE images; 3) for each dataset, the same LB-FCN *light* classifier was trained solely on the synthetic normal and abnormal images, and tested, exclusively on the respective real images.

Aiming to a fair comparison between the classification performance results obtained using the KID and Kvasir-Capsule datasets, the same number and proportion of normal and abnormal images was considered (*i.e.*, 728 normal and 227 abnormal synthetic images). In all experiments, a stratified 10-fold cross validation approach was adopted to alleviate a potential selection bias. More specifically, both the synthetic dataset and the dataset with the real images were split into ten subsets that were fully disjoint, from which, nine subsets of the synthetic dataset were used for training, and a subset from the real dataset was used for testing. The process was repeated ten times, selecting different training and testing subsets, until all the real subsets were used for testing. The training settings of LB-FCN *light* are the ones suggested in [48]. Considering that the generation of the synthetic images was random, the whole experimental procedure was also repeated ten times and average results with standard deviations were recorded.

TABLE 1. Classification results (AUC %) on real images using either real or solely synthetic training images.

	KID [54]	*Kvasir-Capsule* [55]
Real	90.9 ± 0.8	80.0 ± 0.7
TS-GAN	79.1 ± 0.7	68.8 ± 0.7
CycleGAN	62.9 ± 1.2	72.1 ± 1.1
StyleGANv2	61.7 ± 1.1	71.8 ± 1.0
EndoVAE	81.9 ± 0.9	71.3 ± 0.3
TIDE	**89.4 ± 1.2**	**80.2 ± 0.6**

The results of the quantitative experimental evaluation of the TIDE model, were compared with the results obtained by relevant state-of-the-art models for WCE image generation, *i.e.*, TS-GAN [47], and EndoVAE [48]. The respective classification performances are summarized in Table 1, which also includes the reference results obtained per dataset using the real images. It can be noticed that TIDE offers an improved classification performance over all the compared models, and more importantly, it is comparable to the reference one. This validates the hypothesis that real training images can be substituted by synthetic ones, without sacrificing classification performance.

It should also be noted that we have also experimented using methods that have been previously applied for the generation of FE images, including CycleGAN [32] and StyleGANv2 [13]. Training these architectures was challenging, mainly because of problems deriving from the small number of the available WCE training samples. Such issues include, low-quality image generation, lack of diversity, presence of artifacts and more importantly mode-collapse [63]. Initially we tried to take advantage of pre-trained models (on ImageNet [7]) for weight initialization. However, both networks could not converge, resulting in mode collapse, early on training. Training CycleGAN and StyleGANv2 from scratch resulted in images that, while in some cases resembled WCE images, they were mostly unnatural, with noise artifacts being prevalent. This is reflected in the lower classification performance obtained using the datasets generated by CycleGAN and StyleGANv2, as indicated in Table 1.

Aiming to further enhance the classification performance we have also investigated the aspect of training the TIDE architecture on a joint dataset composed of both KID and Kvasir-Capsule datasets. TIDE architecture was trained once on the normal joint subsets of the two datasets, and then it was retrained on the abnormal joint subsets. Then, the synthetic images generated from TIDE were used for training the LB-FCN *light* classifier, which was tested on the real datasets, by following the same experimental procedure described in this section. However, the synthetic images generated after this joint training of TIDE architecture significantly degraded the classification performance of the LB-FCN classifier on the KID and Kvasir-Capsule datasets, which in terms of AUC it was 69.3% ± 3.1% and 73.9% ± 3.9%, respectively. This is likely due to the differences between the two real training datasets in the representation of tissues, *e.g.*, in color and texture, since they have been acquired with different capsule endoscopes.

C. QUALITATIVE EVALUATION

A qualitative, visual, comparison between the real images of the KID and Kvasir-Capsule datasets, and representative images generated by TIDE, can be performed by examining the images of Fig. 3 and Fig. 4, respectively. Real normal images from the small bowel illustrate healthy mucosa which presents circular folds attributed to the folds of the lumen. The texture of the real endoscopic tissue varies due to the existence of intestinal villi. Real abnormal images contain inflammatory lesions which are flat or excavated erosions on the surface of the mucosa characterized by soft or more intense color gradations and sometimes they are covered by a tiny fibrin layer. Figures 3 and 4 show that TIDE generates realistic endoscopic images with a diversity resembling that of the real images. Particularly, it can be noticed that in the synthetic images generated by TIDE, the visible characteristics of the real tissues are preserved, including color, texture and shape. The lesions generated in the case of the abnormal images, not only look like the ones in the real abnormal images, but they are also naturally positioned and blended with the normal tissue. Furthermore, the generated images include realistically reproduced bubbles and debris, which are common in real WCE images.

Figures 5 and 6 provide a comparative visualization of normal and abnormal images, respectively, generated using different state-of-the-art generative models, namely TS-GAN [47], CycleGAN [32], StyleGANv2 [13], the hybrid

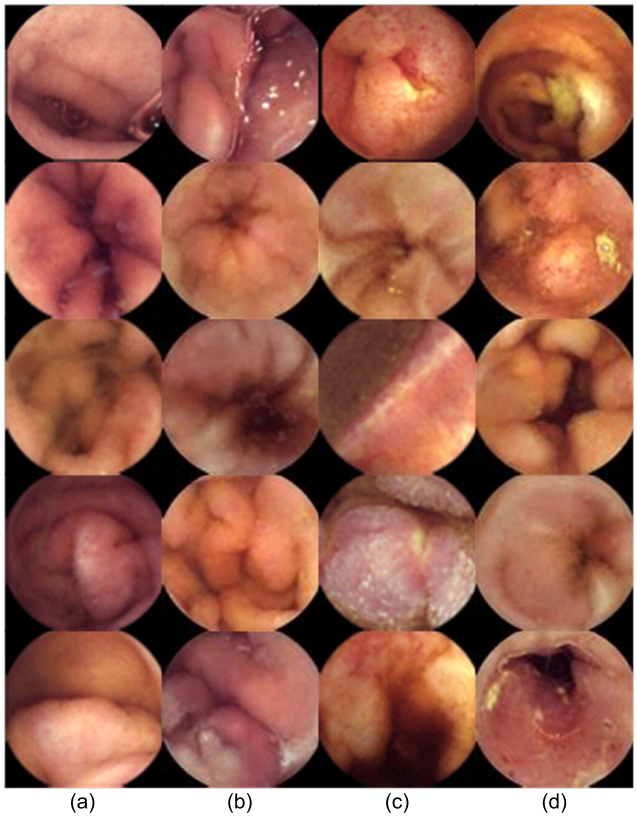


FIGURE 3. Real and synthetic endoscopy images illustrating small bowel tissue from KID dataset. (a) Real normal images. (b) Normal images generated by TIDE. (c) Real abnormal images. (d) Abnormal images generated by TIDE.

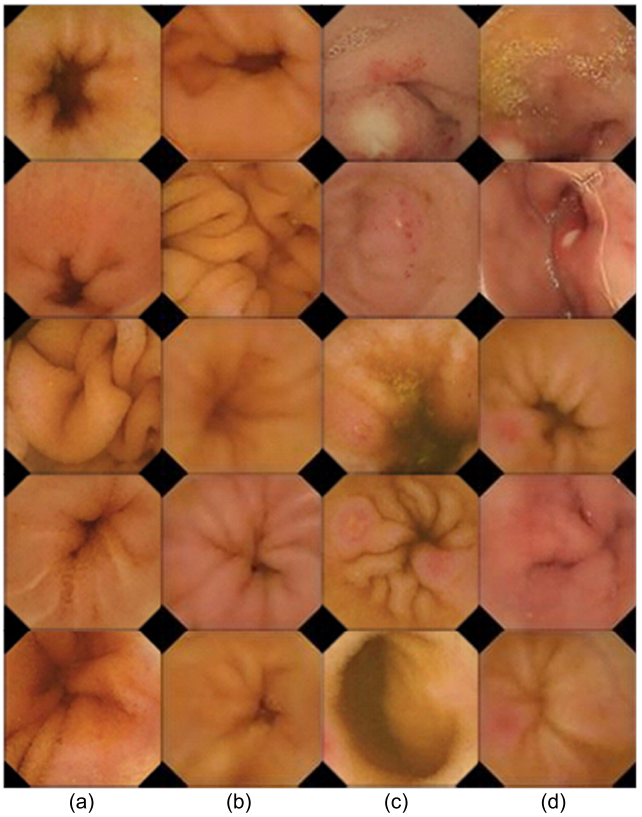


FIGURE 4. Real and synthetic endoscopy images illustrating small bowel tissue from Kvasir dataset. (a) Real normal images. (b) Normal images generated by TIDE. (c) Real abnormal images. (d) Abnormal images generated by TIDE.

VAE/GAN model proposed in [46], UNIT VAE-GAN [36], and EndoVAE [48]. More specifically, with respect to the synthetic normal images in Fig. 5, from a medical viewpoint, TS-GAN (Fig. 5(a)) provides rather realistic-looking synthetic images but with a marked granularity of the image and pixelation that is not common in the usual small-bowel capsule endoscopy images; CycleGAN (Fig. 5(b)), provides an entirely pixelated set of images that, together, the degree of haziness, does not allow any safe observations to be performed with that set; the images generated by StyleGANv2 (Fig. 5(c)) compared with CycleGAN, represent a much-improved version but still suffer from marked image haziness and an outcome that points toward a non-realistic set of normal small bowel images; the hybrid VAE/GAN model (Fig. 5(d)) synthesizes pixelated normal small bowel images with color artifacts that fail to capture the internal circular folds of the small bowel intestine; UNIT VAE-GAN (Fig. 5(e)) provides a blurry set of images in which the anatomy of the small bowel is not realistically reproducible. Also, it can be noticed that UNIT VAE-GAN modifies the boundaries of the images. EndoVAE (Fig. 5(f)) generated a set of realistic normal small-bowel images, which, despite the marked improvement as compared with the results of the previous models (even with TS-GAN), the amount of haziness and the presence of ultra-white artefacts give it away as a non-real dataset; the images generated by TIDE (Fig. 5(g)) are characterized by clarity and higher definition as compared to the previous ones, with only scarce presence of artifacts; the additional water/air bubble interface helps in providing extra realistic features.

Regarding the abnormal images in Fig. 6, all GAN-based (Fig. 6(a-c)) and hybrid GAN-VAE based (Fig. 6(d-e)) methods show non-realistic abnormalities of the small bowel. Although TS-GAN is an early model for WCE image generation, it provides images with a rather clear impression of possible mucosal infiltration/induration by a relevant process. However, the images lack clarity, including some artifacts, and they cannot be used to deduct diagnostic conclusions. The UNIT VAE-GAN model provides images with inflamed tissues, but, despite that, they suffer from marked haziness and blurriness which expose their artificial origin. On the contrary, the last column includes images of mucosal ulceration, characterized by a realistic texture that approximates that of real images and they can be used for clinical training and other functions.

To validate the visual observations with respect to the diversity of the generated images a complementary experimental study was conducted. The exponential of the Shannon entropy of the eigenvalues of a kernel similarity matrix was considered as a generic domain-independent

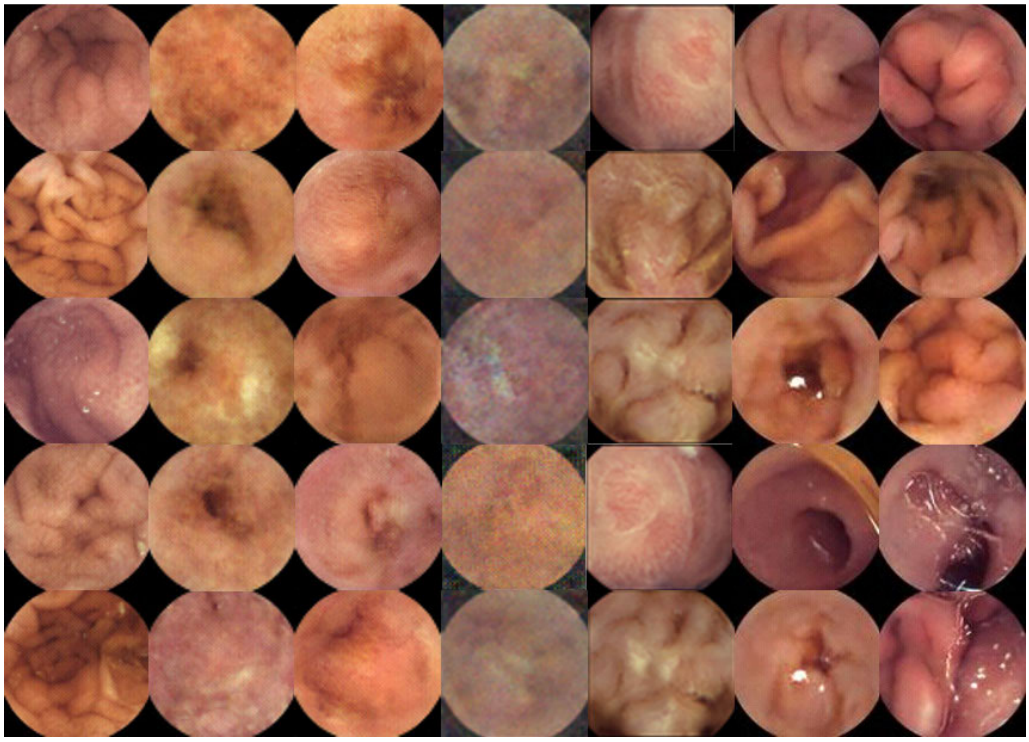


FIGURE 5. Synthetic normal WCE images produced by different generative models. (a) TS-GAN [47] (b) CycleGAN [32] (c) StyleGANv2 [13] (d) hybrid VAE/GAN [46] (e) UNIT VAE-GAN [36] (f) EndoVAE [48] (g) TIDE.

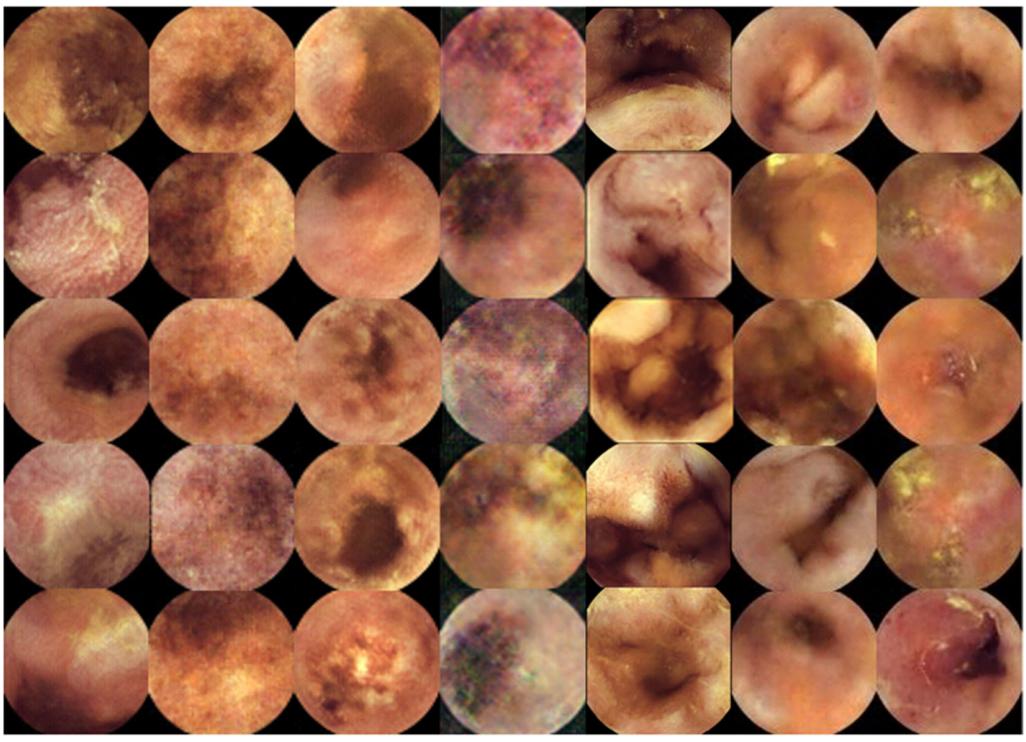


FIGURE 6. Synthetic abnormal WCE images produced by different generative models. (a) TS-GAN [47] (b) CycleGAN [32] (c) StyleGANv2 [13] (d) hybrid VAE/GAN [46] (e) UNIT VAE-GAN [36] (f) EndoVAE [48] (g) TIDE.

measure [64]. Let us consider a collection of independent samples $x_1, x_2, ..., x_n \in \mathcal{X}$, $\mathbf{K} \in \mathbb{R}^{n \times n}$ a positive semi-definite kernel similarity matrix with $\mathrm{K}_{i,i} = 1 \; for \; i \in \{1, \ldots n\}$, and $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$ a vector with the eigenvalues of $\mathbf{K}/n$.

The diversity of this collection of samples can be defined as:

$$\delta = exp\left(-\sum_{i=1}^{S} \lambda_i \log \lambda_i\right) \tag{5}$$

Different kernel similarity matrices can be utilized to capture the visual or semantic similarity of the samples to be evaluated. In this work both pixel-based and feature-based similarity kernels were considered, aiming to quantify the diversity with respect to the image details and semantic content, respectively. Pixel-based similarity is measured as the cosine similarity between pixel vectors, and it captures differences related to low-level image features, such as the brightness and color of the images compared. The feature-based similarity is calculated as the cosine similarity between high-level features of the images. The last pooling layer of an Inception-v3 model trained on ImageNet was selected as a perceptually-relevant feature extractor, which has also been effectively applied in the context of WCE image representation [65]. Additionally, an indicative texture similarity kernel based on Local Binary Patterns (LBP) [66] was selected to assess the degree to which the textural features of the original images is preserved. The texture-based similarity was calculated as the cosine similarity between the texture vectors extracted as in [64]. Ultimately, the diversity δ_g of a generated dataset should be approximately equal with the diversity δ_r of the respective real dataset, *i.e.*, $\delta_{g=}\delta_r$. We consider the relative diversity, as a more meaningful measure defined as $\tilde{\delta} = \delta_g / \delta_r$, because it provides a diversity score that is independent from the diversity of the real dataset used to train the generative model, enabling direct comparisons among different datasets; therefore, this measure is maximized for $\delta_{g=}\delta_r$, *i.e.*, $\tilde{\delta} = 1$. Figure 7 illustrates the relative diversity over all the datasets generated in this study. Considering the feature-based similarities, in that figure it can be noticed that those generated by the two VAE-based models and TS-GAN have higher relative diversities than CycleGAN and StyleGANv2. The improved performance of TS-GAN over StyleGANv2 and CycleGAN is possibly due to the patch-based synthesis it employs, that favors the learning of texture patterns of endoscopic images. Considering the pixel similarities, the results of all models are comparable to each other, except from the abnormal dataset generated by TS-GAN, which exhibits a significantly higher relative diversity. However, TIDE outperforms EndoVAE with respect to the feature-based diversity observed in the normal datasets. Considering the texture-based similarity, from the results presented in Fig. 7 it can be inferred that the VAE-based models, *i.e.*, TIDE and EndoVAE, can reproduce the textural features of the endoscopic tissue with a higher fidelity than the GAN-based methodologies.

Figure 8 presents indicative examples of the images generated by TIDE when trained on the joint dataset composed of both KID and Kvasir-Capsule datasets. As it can be observed, TIDE synthesizes images that are vaguer with a somewhat

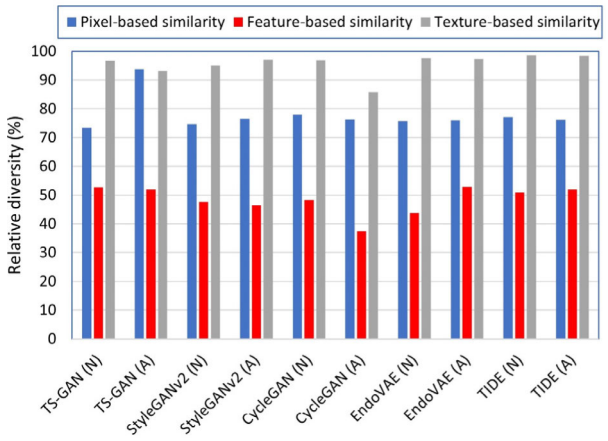


FIGURE 7. Relative diversity, based on pixel feature and texture similarity kernels, of the (N)ormal and the (A)bnormal samples of all the synthetic datasets produced by different generative models.

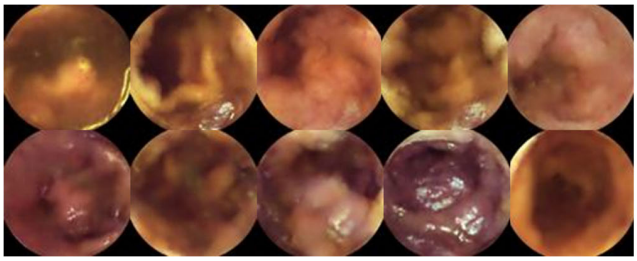


FIGURE 8. Normal (first row) and inflammatory (second row) images generated by TIDE architecture when trained on the joint dataset composed of both KID and Kvasir-Capsule datasets.

distorted content, indicating that the texture of the original endoscopic images is not sufficiently preserved.

Another qualitative study was performed to investigate if the detection of abnormalities in synthetic and real test images is based on similar cues, since this would provide additional evidence on their resemblance. To this end, a well-recognized model-agnostic post-hoc methodology enabling the interpretation of a classifier's outcome, called Grad-CAM [67], was employed. This methodology creates heatmap visualizations highlighting the areas with the higher influence on a machine learning model's prediction. A WCE specialist visually validated that the interpretations obtained from the application of Grad-CAM on the LB-FCN *light* classifier (trained on synthetic images as described in section IV-B) focus on the inflammatory lesions, *i.e.*, the heatmaps are overlapping with the lesions, in both the synthetic and real KID and Kvasir-Capsule datasets. Therefore, also from the perspective of machine learning interpretability the synthetic images generated by TIDE resemble the real ones. Furthermore, the focus of the interpretations on the inflammatory lesions also in the synthetic images, indicates the preservation of the relevant clinical characteristics of the real lesions. Indicative Grad-CAM interpretations are illustrated in Figs. 9-12. These figures show representative heatmaps interpreting the classification of the abnormal images of Fig. 3(c), 3(d) and Fig. 4(c), 4(d) respectively.

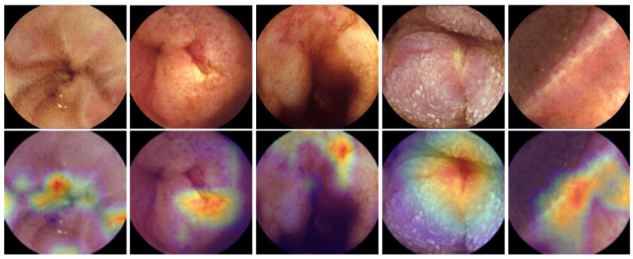


FIGURE 9. Real WCE images with inflammatory conditions from the KID dataset (first row) and their corresponding heatmaps (second row) generated using the Grad-CAM [67] methodology on the trained LB-FCN light classifier [61].

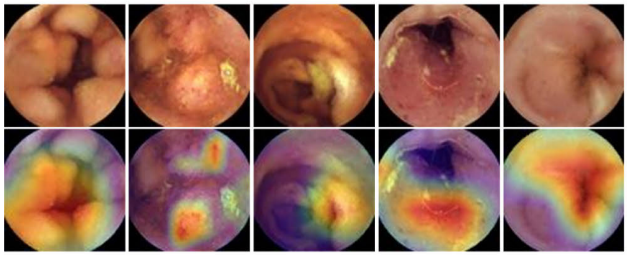


FIGURE 10. Synthetic WCE images with inflammatory conditions from the KID dataset (first row) and their corresponding heatmaps (second row) generated using the Grad-CAM [67] methodology on the trained LB-FCN light classifier [61].

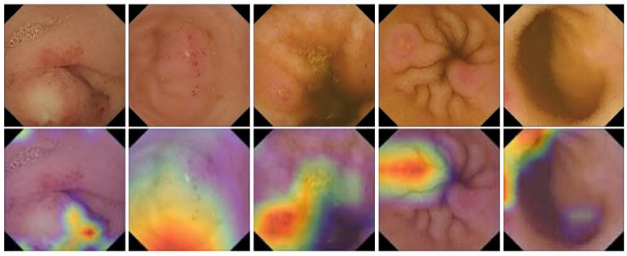


FIGURE 11. Real WCE images with inflammatory conditions from the Kvasir-Capsule dataset (first row) and their corresponding heatmaps (second row) generated using the Grad-CAM [67] methodology on the trained LB-FCN light classifier [61].

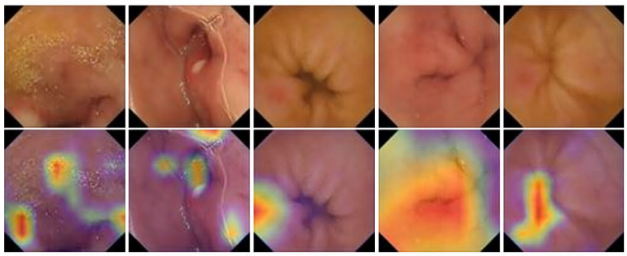


FIGURE 12. Synthetic WCE images with inflammatory conditions from the Kvasir-Capsule dataset (first row) and their corresponding heatmaps (second row) generated using the Grad-CAM [67] methodology on the trained LB-FCN light classifier [61].

It can be noticed that in most cases the higher activation areas (yellow/red colored) of the heatmap overlap with the inflammatory lesions.

D. USER EVALUATION

Based on the above, the datasets generated by TIDE result in a classification performance that is equivalent to that obtained using the respective real WCE images. Also, the images of the TIDE datasets have both a more realistic appearance and they are clearer than the datasets produced by the compared generative models. To investigate if the TIDE datasets are sufficiently realistic also for endoscopists specialized in WCE, a series of Visual Turing Tests (VTTs) was conducted. More specifically, three tests were performed using normal images and images depicting inflammatory conditions from the small bowel, automatically generated by TIDE. The first and second tests included 55 images each, with the first one having only real images from the KID database and the second one containing only synthetic ones generated by the TIDE architecture. The last test combined the two, including 110 images in total. The tests were then given to three endoscopists with 10 to 30 years of experience who were called to distinguish the synthetic from the real images. It is important to note that to avoid any selection bias, the proportion of the real and fake images in the all the tests was undisclosed to the participated experts. Informed consent was obtained and the outcomes of each VTT were not announced to them until the completion of this study.

Table 2 summarizes the results of all the VTTs conducted by the three endoscopists. In the first VTT consisting of only artificially generated images produced by the TIDE architecture, the average accuracy obtained by the endoscopists was 46.1±7.3% (ranging between 38.2% and 52.7%). For the second VTT, which consists of only real images, the mean accuracy was 66.1±8.6% (ranging between 56.4% and 72.7%). The third VTT contained real and artificially synthesized images generated by the proposed methodology; the average obtained accuracy was 50.0±1.8% (ranging between 48.2% and 51.8%). Considering the real images as positive predictions and the synthetic ones as negative predictions, the mean sensitivity and specificity were 65.5±10.1% (ranging between 56.4% and 76.4%) and 34.6±6.5% (ranging between 27.3% and 40.0%), respectively (Table 3). The above results validate that the endoscopic images generated by TIDE are hard to distinguish from the real ones.

TABLE 2. Summary of all visual turing test results.

	Visual Turing Tests		
	1st VTT	2nd VTT	3rd VTT
	Accuracy (%)	*Accuracy (%)*	*Accuracy (%)*
Endoscopist I	38.2	72.7	50.0
Endoscopist II	47.3	69.1	51.8
Endoscopist III	52.7	56.4	48.2
Mean	46.1 ± 7.3	66.1 ± 8.6	50.0 ± 1.8

V. DISCUSSION

The limited availability of annotated datasets in medical imaging is a barrier to essential progress in developing

TABLE 3. Mean sensitivity and specificity in the case of the 3rd VTT.

	3rd VTT	
	Specificity (%)	*Specificity (%)*
Endoscopist I	63.6	36.4
Endoscopist II	76.4	27.3
Endoscopist III	56.4	40.0
Mean	*65.5 ± 10.1*	*34.6 ± 6.5*

image-based CDS systems. Particularly in the domain of WCE the need for such progress is urgent, as the diagnostic yield remains low, and WCE specialists reach their limits by trying to maintain their concentration undistracted while examining several thousands of images [2]. This study was motivated by the need for publicly available benchmarking WCE datasets that will trigger productive competition among image analysis researchers to effectively improve their methods for use in clinical practice. We proposed a multiscale residual VAE architecture capable of generating synthetic WCE images, and showed, using publicly available datasets, that such images can replace the real ones for training machine learning systems for the detection of abnormalities. The range of abnormalities that can be found in the small bowel, where WCE is mainly applicable, is broad. As a proof-of-concept, this study focused on inflammatory lesions, which represent a range of abnormalities associated with diseases, such as IBD and CD, affecting millions of individuals worldwide [1].

During the last decade, image generation methods have been proposed in various medical and non-medical domains. GANs and their variants have had a tremendous success mainly in generating synthetic images of human faces, and several studies have reported exceptional results in the generation of medical images [26], [28], [33], [34]. However, important factors of success in these studies constitute the large number of diverse training data, and the relatively aligned content of the training images, *e.g.*, the face images are aligned with respect to the facial features, and CT or MRI images can be aligned with respect to the depicted body structures. On the other hand, the generation of synthetic endoscopic images of the GI tract is more challenging, since their content is more diverse without features that could be considered for alignment. Related studies (Section II) have reported results based on significantly larger, usually not publicly available, training sets. These studies have also indicated issues with respect to the application of GAN-based generative models. For example, it is worth noting that [21] reports that a contemporary classification system, trained with images generated by a GAN model, reached saturation in performance improvement after a certain point, even if more synthetic polyp images were added to the training set. This was attributed to the fact that the GAN model was unable to introduce new unseen features. The GAN was only manipulating the existing features in the training set, trying to reuse the same set of features to generate new-looking synthetic polyps. This is a common limitation of GAN-based image generation models, and it could also justify the results of TS-GAN in our study. Although that model managed to generate more plausible images than the other compared GAN models, with a relatively high diversity, the generated dataset was not sufficient to provide a classification performance equivalent to that of the respective real datasets.

The classification performance using synthetic datasets generated from the KID database was generally higher than that observed using synthetic datasets generated from the Kvasir-Capsule dataset, regardless of the type of the generation model. This could be attributed to the fact that the real Kvasir-Capsule images generally include smaller lesions than those of the real KID dataset. It is worth mentioning that in the case of inflammatory image synthesis, there are some cases where TIDE architecture generates images with lesions that are not readily discernible. Although one can consider this as a weakness of the model, this can be partially attributed to the fact that the real datasets include images with inflammatory conditions, *e.g.*, erosions, which are hard even for experts to distinguish. Thus, in some cases TIDE architecture generates variations of these images that may be even harder distinguishable.

The results obtained from the two VAE architectures compared in this study, namely EndoVAE and TIDE, indicate that although they can both generate quite realistic images, the improved diversity, and the higher definition of the depicted structures in the images generated by TIDE, play a significant role in the improvement of the classification performance. The improved performance of TIDE can be attributed to the use of multiscale blocks which enable more diverse features to be encoded and decoded from and to the latent space [52]. This feature diversity gives the opportunity to create a swallower architecture, minimizing the number of free parameters, without compromising the generative performance and as a result enables the network to be trained on smaller datasets [68]. Also, the qualitative results validate that multiscale information can enhance image quality, which has also been observed in the context of multiscale GANs [31]. The fact that multiscale representations can provide more detailed content characterizations' in the context of endoscopic image analysis has also been noted in the recent review study of Ali [69]. In addition, the residual connections introduced in this paper, allow the gradient to be propagated easily throughout the network, further enhancing the quality of the extracted features. Furthermore, the VAE-based learning avoids the adversarial optimization process followed in GAN models, which demands a large amount of data in order to be properly trained [15].

The user-evaluation study validated the clinical relevance of the images generated by the proposed TIDE architecture. Evaluating synthetic medical images in the context of unconditional generation, in which there are no explicit

pairs of ground truth and synthetic data, is still a challenge [70]. Contrary to the assessment of synthetic natural images, *e.g.,* natural scenes and human faces, in the complex context of evaluating synthetic medical images, the issue of domain-specific image quality metrics that capture the clinical relevance of the generated images, has yet to be addressed [71]. A limiting factor is that the existing measures require a rather large reference dataset or distribution over samples in order to be properly estimated [58], [64]. The score used in our study is independent of such a requirement and therefore, it can be applied to any generative model and data domain for a given similarity measure [64].

In practice, TIDE could be used to generate synthetic annotated WCE datasets based on real, anonymized datasets, and the real datasets can be securely kept, within the premises of the healthcare provider. The synthetic datasets can be shared publicly with the (technical) research community, without raising any legal or ethical concerns, since they are not real, originating from statistical processing that does not allow identification of any personal information. A limitation of the proposed methodology is that TIDE generates images belonging to one class, and to generate images belonging to multiple classes, it needs to be trained separately with data from each different class. This is in agreement with the observation made in [72] indicating that deep learning algorithms for WCE tend to be more accurate when trained independently on different classes of abnormalities than on multiclass data. While we tried to divide the latent space so that it can capture more classes, the network was becoming unstable, and the quality of the results was affected. This is possibly due to the nature of endoscopy images which present a high inter-class similarity [73], making it more difficult for the network to distinguish their differences. Efficiently disentangling the latent space of VAE models is still a research challenge, and to the best of our knowledge, there are no relevant studies dedicated to the generation of endoscopy images using VAEs.

This study investigated the post-hoc interpretability of inflammatory lesion detection through visual interpretation of synthetic image classification. The qualitative assessment of the results in comparison to the results obtained using real images provided additional indications on the resemblance of the synthetic with the real images. This is a first approach towards the use of machine learning interpretability in the context of image synthesis. There are still other aspects to this direction that have yet to be further investigated, such as the development of interpretable generative models. Current methodologies for the generation of endoscopic images, including TIDE, are lacking interpretability in that sense, and this could be considered as a limitation for their adoption. The interpretation of endoscopic image synthesis at that level could shed light on the internal processes that lead to the generation of different parts of the synthetic images considering the different properties of tissues, *e.g.*, color and texture, depicted in respective real images; thus, enhancing the clinicians' trust in the generative systems.

VI. CONCLUSION

This paper presented TIDE, a novel VAE architecture for the generation of synthetic WCE images that incorporates multiscale feature extraction and residual learning in the context of variational image synthesis. A proof-of-concept case study was investigated addressing the generation of normal and abnormal images of the small bowel in the context of image-based CDS systems for the detection of inflammatory small bowel lesions. More specifically, in the study conducted, the performance of the proposed TIDE architecture was evaluated quantitatively and qualitatively following similar experimental procedures used in relevant studies presented in Section II. All the experimental evaluation was performed on publicly available endoscopic datasets. For the quantitative assessment, the synthetic image datasets produced by the proposed TIDE architecture were solely employed to train a state-of-art classifier for the recognition of inflammatory conditions. The performance of the trained classifier was then evaluated on real image datasets. Apart from TIDE, various generative models, such as VAEs, GANs and hybrid VAE/GAN models, were tested in endoscopic image synthesis containing inflammation conditions. Hybrid VAE/GAN based methodologies have not previously been evaluated in that context. Furthermore, the synthesis performance of the proposed TIDE architecture was investigated after being trained on a joint dataset. A qualitative comparison was performed to evaluate the diversity of the synthetic datasets produced, and a user evaluation study was conducted to assess the clinical relevance of the images generated by the proposed architecture. The results of the experimental evaluation of TIDE lead to the following main conclusions about the proposed architecture:

- It enables the generation of synthetic images of enhanced clarity and diversity, suitable to fully substitute real training sets for WCE image classification.
- It accomplishes effective and realistic WCE image synthesis even using a limited number of training samples.
- The synthetic images generated by TIDE are difficult to distinguish even by experienced WCE specialists.

Future research directions include the application of the proposed framework for generating images from the entire GI tract with various abnormalities and pathological conditions. To this direction, we are planning to extend the TIDE architecture to enable multi-class training and investigate methods for joint/cross-dataset image synthesis using VAEs. Furthermore, the generality of the proposed architecture makes it a candidate solution for generating synthetic images of other medical imaging modalities. Finally, a promising direction is the investigation of interpretable/explainable generative models and their application to GI endoscopy.

ACKNOWLEDGMENT

The authors would like to thank the anonymous endoscopists who participated in their study.

REFERENCES

[1] A. D. Sperber et al., "Worldwide prevalence and burden of functional gastrointestinal disorders, results of Rome foundation global study," *Gastroenterology*, vol. 160, no. 1, pp. 99–114, 2021.

[2] X. Dray, D. Iakovidis, C. Houdeville, R. Jover, D. Diamantis, A. Histace, and A. Koulaouzidis, "Artificial intelligence in small bowel capsule endoscopy-current status, challenges and future promise," *J. Gastroenterol. Hepatol.*, vol. 36, no. 1, pp. 12–19, Jan. 2021.

[3] P. Handa, N. Goel, S. Indu, and D. Gunjan, "Automatic detection of colorectal polyps with mixed convolutions and its occlusion testing," *Neural Comput. Appl.*, vol. 35, no. 26, pp. 19409–19426, Sep. 2023.

[4] N. Goel, S. Kaur, D. Gunjan, and S. J. Mahapatra, "Dilated CNN for abnormality detection in wireless capsule endoscopy images," *Soft Comput.*, vol. 26, no. 3, pp. 1231–1247, Feb. 2022.

[5] Palak, H. Mangotra, and N. Goel, "Effect of selection bias on automatic colonoscopy polyp detection," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104915.

[6] N. Goel, S. Kaur, D. Gunjan, and S. J. Mahapatra, "Investigating the significance of color space for abnormality detection in wireless capsule endoscopy images," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103624.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

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

[9] G. Pascual, P. Laiz, A. García, H. Wenzek, J. Vitrià, and S. Seguí, "Time-based self-supervised learning for wireless capsule endoscopy," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105631.

[10] 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," *Sci. Rep.*, vol. 12, no. 1, p. 261, Jan. 2022.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[14] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[15] D. Saxena and J. Cao, "Generative adversarial networks (GANs): Challenges, solutions, and future directions," *ACM Comput. Surveys*, vol. 54, no. 3, pp. 1–42, May 2021.

[16] 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," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1887–1891.

[17] 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," in *Proc. Med. Image Comput. Comput. Assist. Intervent.-MICCAI 2019: 22nd Int. Conf.*, Shenzhen, China, Oct. 2019, pp. 530–538.

[18] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2572–2581, Dec. 2018.

[19] 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, pp. 56007–56017, 2018.

[20] A. Vats, M. Pedersen, A. Mohammed, and Ø. Hovde, "Evaluating clinical diversity and plausibility of synthetic capsule endoscopic images," *Sci. Rep.*, vol. 13, no. 1, p. 10857, Jul. 2023.

[21] H. A. Qadir, I. Balasingham, and Y. Shin, "Simple U-net based synthetic polyp image generation: Polyp to negative and negative to polyp," *Biomed. Signal Process. Control*, vol. 74, Apr. 2022, Art. no. 103491.

[22] A. Sams and H. H. Shomee, "GAN-based realistic gastrointestinal polyp image synthesis," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–4.

[23] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019, doi: 10.1561/2200000056.

[24] J. Ahn, H. Nguyen Loc, R. Krishna Balan, Y. Lee, and J. Ko, "Finding small-bowel lesions: Challenges in endoscopy-image-based learning systems," *Computer*, vol. 51, no. 5, pp. 68–76, May 2018.

[25] D. E. Diamantis and D. K. Iakovidis, "ASML: Algorithm-agnostic architecture for scalable machine learning," *IEEE Access*, vol. 9, pp. 51970–51982, 2021.

[26] C. Bermudez, A. J. Plassard, L. T. Davis, A. T. Newton, S. M. Resnick, and B. A. Landman, "Learning implicit brain MRI manifolds with deep learning," *Med. Imag. Image Process.*, vol. 10574, pp. 408–414, Mar. 2018.

[27] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2018, pp. 580–588.

[28] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, "CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2018, pp. 732–740.

[29] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.

[30] C. Baur, S. Albarqouni, and N. Navab, "Generating highly realistic images of skin lesions with GANs," in *Proc. OR 2.0 Context-Aware Operating Theaters, Comput. Assist. Robotic Endoscopy, Clin. Image-Based Procedures Skin Image Anal., 1st Int. Workshop*, 2018, pp. 260–267.

[31] B. Zhan, D. Li, X. Wu, J. Zhou, and Y. Wang, "Multi-modal MRI image synthesis via GAN with multi-scale gate mergence," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 17–26, Jan. 2022.

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[33] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, Z. Xu, and J. Prince, "Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN," in *Proc. Deep Learning Med. Image Anal. Multimodal Learn. Clin. Decis. Support, 4th Int. Workshop*. Cham, Switzerland: Springer, 2018, pp. 174–182.

[34] J. Cai, Z. Zhang, L. Cui, Y. Zheng, and L. Yang, "Towards cross-modal organ translation and segmentation: A cycle- and shape-consistent generative adversarial network," *Med. Image Anal.*, vol. 52, pp. 174–184, Feb. 2019.

[35] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abramoff, A. M. Mendonca, and A. Campilho, "End-to-end adversarial retinal image synthesis," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 781–791, Mar. 2018, doi: 10.1109/TMI.2017.2759102.

[36] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[37] A. U. Hirte, M. Platscher, T. Joyce, J. J. Heit, E. Tranvinh, and C. Federau, "Realistic generation of diffusion-weighted magnetic resonance brain images with deep generative models," *Magn. Reson. Imag.*, vol. 81, pp. 60–66, Sep. 2021.

[38] H. Huang, R. He, Z. Sun, and T. Tan, "Introvae: Introspective variational autoencoders for photographic image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

[39] F. Gao, T. Wu, X. Chu, H. Yoon, Y. Xu, and B. Patel, "Deep residual inception encoder–decoder network for medical imaging synthesis," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 39–49, Jan. 2020.

[40] D. K. Iakovidis et al., "Roadmap on signal processing for next generation measurement systems," *Meas. Sci. Technol.*, vol. 33, no. 1, 2021, Art. no. 012002.

[41] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, May 2019, pp. 4570–4580.

[42] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12104–12114.

[43] L. Sharan, G. Romano, S. Koehler, H. Kelm, M. Karck, R. De Simone, and S. Engelhardt, "Mutually improved endoscopic image synthesis and landmark detection in unpaired image-to-image translation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 127–138, Jan. 2022.

[44] M. Pennazio et al., "Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European society of gastrointestinal endoscopy (ESGE) guideline-update 2022," *Endoscopy*, vol. 55, no. 1, pp. 58–95, 2022.

[45] E. Rondonotti, A. Koulaouzidis, D. E. Yung, S. N. Reddy, J. Georgiou, and M. Pennazio, "Neoplastic diseases of the small bowel," *Gastrointestinal Endoscopy Clinics*, vol. 27, no. 1, pp. 93–112, 2017.

[46] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, Feb. 2016, pp. 1558–1566.

[47] 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," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2019, pp. 519–524.

[48] D. E. Diamantis, P. Gatoula, and D. K. Iakovidis, "EndoVAE: Generating endoscopic images with a variational autoencoder," in *Proc. IEEE 14th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jun. 2022, pp. 1–5.

[49] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[50] J. Li and D. Liu, "Information bottleneck theory on convolutional neural networks," *Neural Process. Lett.*, vol. 53, no. 2, pp. 1385–1400, Apr. 2021.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[53] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[54] A. Koulaouzidis, D. Iakovidis, D. Yung, E. Rondonotti, U. Kopylov, J. Plevris, E. Toth, A. Eliakim,G. W. Johansson, W. Marlicz, G. Mavrogenis, A. Nemeth, H. Thorlacius, and G. Tontini, "KID project: An Internet-based digital video atlas of capsule endoscopy for research purposes," *Endoscopy Int. Open*, vol. 5, no. 6, pp. E477–E483, Jun. 2017.

[55] P. H. Smedsrud et al., "KVASIR-Capsule, a video capsule endoscopy dataset," *Sci. Data*, vol. 8, no. 1, p. 142, 2021.

[56] D. K. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Comput. Med. Imag. Graph.*, vol. 34, no. 6, pp. 471–478, Sep. 2010.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent., ICLR*, San Diego, CA, USA, 2015, pp. 1–15.

[58] A. Borji, "Pros and cons of GAN evaluation measures: New developments," *Comput. Vis. Image Understand.*, vol. 215, Jan. 2022, Art. no. 103329.

[59] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.

[60] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, 2016.

[61] D. E. Diamantis, D.-C. C. Koutsiou, and D. K. Iakovidis, "Staircase detection using a lightweight look-behind fully convolutional neural network," in *Eng. Appl. Neural Networks*, J. Macintyre, L. Iliadis, I. Maglogiannis, and C. Jayne, Eds. Cham: Springer, 2019, pp. 522–532.

[62] 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, pp. 192–201, Mar. 2019.

[63] H. Thanh-Tung and T. Tran, "Catastrophic forgetting and mode collapse in GANs," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–10.

[64] D. Friedman and A. B. Dieng, "The vendi score: A diversity evaluation metric for machine learning," *Trans. Mach. Learn. Res.*, 2023. [Online]. Available: https://openreview.net/forum?id=g97OHbQyk1

[65] X. Li, H. Zhang, X. Zhang, H. Liu, and G. Xie, "Exploring transfer learning for gastrointestinal bleeding detection on small-size imbalanced endoscopy images," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 1994–1997.

[66] X. Dong, J. Dong, and M. J. Chantler, "Perceptual texture similarity estimation: An evaluation of computational features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2429–2448, Jul. 2021.

[67] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[68] L. Brigato and L. Iocchi, "A close look at deep learning with small data," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2490–2497.

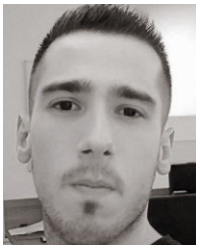
[69] S. Ali, "Where do we stand in AI for endoscopic image analysis? Deciphering gaps and future directions," *npj Digit. Med.*, vol. 5, no. 1, p. 184, Dec. 2022.

[70] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "GANs for medical image analysis," *Artif. Intell. Med.*, vol. 109, Sep. 2020, Art. no. 101938.

[71] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101552.

[72] S. Pecere, M. F. Chiappetta, L. E. D. Vecchio, E. Despott, X. Dray, A. Koulaouzidis, L. Fuccio, A. Murino, E. Rondonotti, M. Spaander, and C. Spada, "The evolving role of small-bowel capsule endoscopy," *Best Pract. Res. Clin. Gastroenterol.*, vols. 64–65, Aug. 2023, Art. no. 101857.

[73] X. Guo and Y. Yuan, "Semi-supervised WCE image classification with adaptive aggregated attention," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101733.



DIMITRIOS E. DIAMANTIS was born in Athens, Greece, in 1991. He received the B.Sc. degree in computer science from the Department of Computer Engineering, Technological Educational Institute of Central Greece, in 2014, and the M.Sc. degree (Hons.) in computational medicine and biology and the Ph.D. degree in deep learning from the University of Thessaly, Greece, in 2017 and 2021, respectively.

From 2009 to 2023, he was a Senior Software Engineer, a Solution Architect, and a Chief Technical Officer (CTO) in several worldwide enterprises, where he gained significant experience in the enterprise software industry. Since 2016, he has been a member of the Biological Imaging Laboratory, University of Thessaly. His research interests include signal, image and video analysis, intelligent systems, deep learning, software engineering, and applications.

PANAGIOTA GATOULA received the B.S. degree in computer science and biomedical informatics from the University of Thessaly, Greece, in 2021, where she is currently pursuing the Ph.D. degree with the School of Sciences. Her research interest includes deep generative methodologies.

ANASTASIOS KOULAOUZIDIS received the M.D. degree from the Medical School, Aristotle University of Thessaloniki, Greece, in 1995, the Doctorate in Medicine (D.M.) degree from The University of Edinburgh, in 2014, and the D.Phil. degree from Lund University, in 2020.

He was a Professor with the Department of Clinical Research, Southern University of Denmark (SDU), and Overlaege with the Department of Surgery, Odense University Hospital and Svendborg Hospital, Svendborg, Denmark. Since 2020, he has been an Honorary Professor with the Department of Social Medicine and Public Health, Pomeranian Medical University, Poland. Previously, he has worked for more than 12 years with the Centre of Liver & Digestive Disorders, The Royal Infirmary of Edinburgh, being the Clinical Lead of the capsule endoscopy (CE) service, and an Honorary Clinical Fellow of the School of Clinical Sciences, The University of Edinburgh. He is the coauthor of ten book chapters and more than 280 PubMed-listed publications, out of which more than 160 are in CE (clinical and developmental fields). His research interests include the clinical applications of CE, quality improvement and software diagnostics, minimally invasive endoscopy, and hardware and concept development in CE.

Dr. Koulaouzidis is a member of the editorial and/or advisory board of several specialty journals. He became a member of the Royal College of Physicians of Edinburgh, U.K., in 2004, and ascended to the Fellowship of the same College, in 2013. He is also a fellow of the European Board of Gastroenterology, in 2009, the Royal Society for Public Health, in 2013, the American College of Gastroenterology, in 2015, and the American Society of Gastrointestinal Endoscopy, in 2016. He was awarded the Given® Imaging-ESGE Research Grant, in 2011, a couple of Innovation Initiatives (The University of Edinburgh) Grants, in 2011 and 2016, and an ESGE Postgraduate Visiting Fellow Grant, in 2010. He is also a World Endoscopy Organization (WEO) Star and he has participated in several European Society for Gastrointestinal Endoscopy (ESGE) guidelines. In 2022, he was awarded the Best Senior Researcher Award from SDU. He is an Associate Editor or the Editor-in-Chief of at least three Gastroenterology/Hepatology journals.

DIMITRIS K. IAKOVIDIS (Senior Member, IEEE) received the Ph.D. degree in informatics from the University of Athens, Greece, in 2004.

In 2015, he was appointed as an Associate Professor with the Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece. Currently, he is a Professor, he serves as the Director of the Biomedical Imaging Laboratory, and the Deputy Head of the Department of Computer Science and Biomedical Informatics. His research interests include signal and image processing, decision support systems, intelligent systems, and applications. In this context, he has coauthored over 200 papers in international journals, conferences, and books. He has served as an Associate Editor for IEEE TRANSACTIONS ON FUZZY SYSTEMS and *IET Signal Processing*. He is an Editorial Board Member of *Measurement Science and Technology* and *Sensors* journals.