Semi-Supervised Bladder Tissue Classification in Multi-Domain Endoscopic Images

Jorge F. Lazo, Benoit Rosa, Michele Catellani, Matteo Fontana, Francesco A. Mistretta, Gennaro Musi, Ottavio de Cobelli, Michel de Mathelin, *Senior Member, IEEE*, and Elena De Momi, *Senior Member, IEEE*

Abstract—Objective: **Accurate visual classification of bladder tissue during Trans-Urethral Resection of Bladder Tumor (TURBT) procedures is essential to improve early cancer diagnosis and treatment. During TURBT interventions, White Light Imaging (WLI) and Narrow Band Imaging (NBI) techniques are used for lesion detection. Each imaging technique provides diverse visual information that allows clinicians to identify and classify cancerous lesions. Computer vision methods that use both imaging techniques could improve endoscopic diagnosis. We address the challenge of tissue classification when annotations are available only in one domain, in our case WLI, and the endoscopic images correspond to an unpaired dataset, i.e. there is no exact equivalent for every image in both NBI and WLI domains.** ***Method:*** **We propose a semi-surprised Generative Adversarial Network (GAN)-based method composed of three main components: a teacher network trained on the labeled WLI data; a cycle-consistency GAN to perform unpaired image-to-image translation, and a multi-input student network. To ensure the quality of the synthetic images generated by the proposed GAN we perform a detailed quantitative, and qualitative analysis with the help of specialists.** ***Conclusion:*** **The overall average classification accuracy, precision, and recall obtained with the proposed method for tissue classification are 0.90, 0.88, and 0.89 respectively, while the same metrics obtained in the unlabeled domain (NBI) are 0.92, 0.64, and 0.94 respectively. The quality of the generated images is reliable enough to deceive specialists. Significance: This study shows the potential of using semi-supervised GAN-based bladder tissue classification when annotations are limited in multi-domain data.**

***Index Terms*—Bladder cancer, semi-supervised learning, generative-adversarial networks, image-to-image translation, tissue classification, multi-domain image classification.**

I. Introduction

URINARY tract cancer comprises different types of lesions ranging from benign tumors to aggressive neoplasms with high mortality. This disease had 164,000 patients reported in 2021 and it is among the top 10 most common cancers worldwide [1]. Muscle Invasive Bladder Cancer originates on the inner surface of the bladder and is more likely to metastasize than Non-Muscle Invasive Bladder Cancer (NMIBC) [2]. The gold standard for Bladder Cancer (BC) diagnosis is cystoscopy. In case of finding abnormal tissue, patients should undergo Trans-Urethral Resection of the Bladder Tumor (TURBT) [3]. This procedure consists of the insertion of an endoscope in the urinary tract and the removal of visible tumor lesions.

The World Health Organization WHO has defined a stratification of urothelial carcinoma accordingly to their propensity of invasion and it can be generalized into two main classes: High-Grade Carcinoma (HGC) and Low-Grade Carcinoma (LGC) [4]. Visual classification of BC is a challenging task. The shapes of lesions either high-grade or low-grade tumors are quite similar in some cases, and the visual difference between healthy tissue and non-tumor lesions is not trivial [2]. In fact, definitive diagnosis, staging, and grading of cancer are only possible after histological analysis of the resected tissue [5].

The use of different imaging techniques other than White Light Imaging (WLI), such as Narrow Band Imaging (NBI) can improve the differentiation of tumorous lesions from normal tissue [6], [7]. Samples of different bladder tissue in both image domains are depicted in Fig. 1. In NBI, a white light source is filtered in two narrow bands of 415 nm and 540 nm. At these wavelengths, the hemoglobin reflection spectra present a global and a local maximum respectively [8]. This increases the contrast between the surface mucosa, the capillaries, and the blood vessels in the submucosa, therefore improving bladder cancer

Manuscript received 29 September 2022; revised 18 January 2023; accepted 14 March 2023. Date of publication 10 April 2023; date of current version 28 September 2023. This work was supported in part by the the European Union's Horizon 2020 Research and Innovation Programme through ATLAS Project under Marie Sködowska-Curie under Grant 813782, and in part by the French State Funds managed by the Agence Nationale de la Recherche through the Investissements d'Avenir Program under Grants ANR-11-LABX-0004 (Labex CAMI) and ANR-10-IAHU-02 (IHU-Strasbourg). *(Corresponding author: Jorge F. Lazo.)*

Jorge F. Lazo is with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy and also with the ICube, UMR 7357, CNRS-Université de Strasbourg, F-67412 Illkirch, France (e-mail: jorgefrancisco.lazo@polimi.it).

Benoit Rosa and Michel de Mathelin are with the ICube, UMR 7357, CNRS-Université de Strasbourg, France.

Michele Catellani is with the Hospital Papa Giovanni XXIII, Italy.

Matteo Fontana is with the European Institute of Oncology (IRCCS), Italy.

Francesco A. Mistretta, Gennaro Musi, and Ottavio de Cobelli are with the European Institute of Oncology (IRCCS), Italy and also with the Department of Oncology and Hematology-Oncology, University of Milan, Italy.

Elena De Momi is with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, Italy and also with the European Institute of Oncology (IRCCS), Italy.

The dataset is available at https://zenodo.org/record/7741476#.ZBQUK7TMJ6k.

This article has supplementary downloadable material available at https://doi.org/10.1109/TBME.2023.3265679, provided by the authors.

Digital Object Identifier 10.1109/TBME.2023.3265679

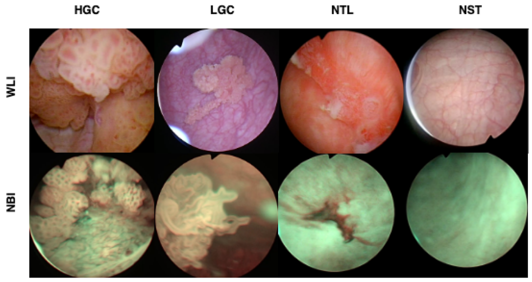


Fig. 1. Sample images of the different classes in the bladder tissue classification dataset. From left to right: High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), No Tumor Lesion (NTL), and Non-Suspicious Tissue (NST).

diagnosis by highlighting visual structures that are hard to notice when using only WLI [9]. Typically during TURBT procedures an initial inspection using WLI is carried out. Subsequently, in a second inspection, the anatomical structures deemed suspicious are examined using NBI to confirm. In some cases, the use of NBI by itself could be more efficient than WLI in the detection of NMIBC [9].

Despite the current advances in optical methods and their implementation in new devices, missing rates are reported to be between 10 and 20% [10]. The clinical interest in endoscopic tissue classification is related to the actions to be performed during surgery, as well as the follow-up treatment. The development of computer-aided diagnosis (CAD) systems for BC classification could help clinicians reduce current miss-classification rates which are related to incomplete excision of tumorous tissue, and cancer recurrence reported to have values of 75% [11]. For example, identifying a high-grade tumor in real-time could lead to the resection of a wider and deeper section of the tissue to avoid future recurrences.

In recent years, Deep-Learning (DL)-based methods have shown promising results in the analysis of endoscopic images. Most of the currently available datasets for endoscopic image analysis focus on colonoscopy [12], [13] and consist mainly of WLI data. Recently, few studies which include NBI data too have stressed on the advantage of using multi-domain data in the colonoscopy scenario [14], [15], [16].

In the case of the urinary system, only a few studies have been carried out in the task of tissue classification from endoscopic images [17], [18], [19], [20]. Except for the study presented in [20] where BL imaging is used, the rest of the studies use only WLI. Multi-domain image classification implies several challenges, especially when the data and annotations are not evenly distributed across the different domains and some of the classes are under-represented [21].

In the specific case of TURBT some of these challenges include the fact that visually it is difficult to differentiate between lesions and the diagnosis is inconclusive [22]. Furthermore, due to the fact that multi-imaging endoscopes can collect only one imaging type at the time, it is not possible to have equivalent pairs of WLI and NBI images. Usually, an initial examination of the bladder is carried out using WLI and the lesions and anatomical structures deemed to be potentially cancerous tissue are examined again with NBI, in case this modality is available which is not always the case. An additional challenge is related to the imbalance of data in terms of the different classes and types of tissue. Non-Suspicious Tissue (NST) usually receives less attention during interventions, therefore fewer amount of image data is collected from it than from lesions, either in WLI or NBI. Furthermore, non-cancerous lesions such as cystitis or other types of bladder inflammations are less common to appear in the initial inspection during TURBT. All this contributes to the fact that most of the datasets (including ours) are imbalanced not only in terms of different image domains but also in terms of tissue classes.

In this work, we focus on the task of bladder tissue classification in multi-domain images from TURBT procedures, with special emphasis in the fact that annotations only exist in one of these image domains. Considering that most state-of-the-art computer vision methods are sensitive to changes in domain [23], and the specific challenges existing in endoscopic image classification, we propose a GAN-based semi-supervised approach which comprises three main components: 1) A teacher network trained on the labeled WLI images. 2) A cycle consistency GAN to perform the unpaired image-to-image translation and 3) A multi-input multi-domain image classifier trained in a semi-supervised way. We show that with our method it is possible to obtain satisfactory classification results even when annotations from one domain are not available.

To ensure that the images produced with the proposed translation network are consistent with the structural and pathological features of the source domain, we perform a detailed quantitative and qualitative analysis of the generative models. Additionally, we validate its quality with help of specialists familiar with the TURBT procedure. In order to allow future research in the task of bladder tissue classification, and ease benchmarking of future methods, we will release the dataset upon publication.

II. Related Work

A. Tissue Classification in Endoscopy

The analysis of endoscopic images has been rapidly developing in recent years thanks to the recent availability of new public datasets [13], [24]. In the specific task of tissue classification different models and techniques have been proposed with a special focus on the gastrointestinal (GI) tract. The existing methods range from the proposal of feature extraction models [25], [26], to the use of transfer learning and pre-trained CNNs [27], [28] and to more complex methods that focus on targeting the specific challenges present when working with GI endoscopic images [29], [30], [31], [32].

In the case of the bladder, Ikeda et al. [19] proposed the use of 2-step transfer learning by first fine-tuning their models on 8728 gastroscopic images, and then re-training the models on 2102 cystoscopy WLI images, using the GoogLeNet model for the task of binary classification of images with and without NMIBC. Yang et al. [18] compared the use of 3 different Convolutional Neural Networks (CNNs) as well as the platform EasyDL. The models used were LeNet, AlexNet and GoogLeNet. Their dataset includes 1200 cystoscopy images with cancer and 1150 without. Shkolyar et al. [17] proposed CystoNet, a CNN for bladder cancer detection and binary classification. In their study, they used 2335 WLI frames of normal benign bladder mucosa and 417 histologically confirmed papillary urothelial carcinoma to train the network. In [33] the use of a Generative Adversarial Network (GAN) is proposed to perform data augmentation, then

AlexNet and VGG16 are trained with the real and augmented data. In total 202 images from a Confocal Laser Endomicroscope were used in their experiments. In [20] Ali et al. proposed the use of pre-trained models for the task of cancer malignancy, grading, and invasiveness classification on BL photodynamic cystoscopy images. The dataset was composed of 261 BL images and the pre-trained models used were VGG16, ResNet-50, MobileNetV2, and InceptionV3. On top of the pre-trained models, a shallow network was added to perform the classification.

B. Image to Image Translation

Since its introduction, GANs have become an outstanding method for different tasks in DL applications. GANs have been used for different purposes on endoscopic images such as the generation of synthetic images to improve polyp detection, or the construction of SLAM models to predict depth maps in colonoscopy [34], [35].

One of the applications of GANs is image-to-image translation. This task can be resumed as the mapping of an image in domain $\mathbb{A}$ to another domain $\mathbb{B}$. In our case, these domains correspond to NBI and WLI. These types of models have been applied in diverse biomedical and endoscopic image tasks such as the translation between optical colonoscopy images and virtual colonoscopy images [36], the mapping between cadaveric and live images [37], the adaptation between phantom images real endoscopic videos among others [38], [39].

Using image-to-image translation with a focus on classification has been previously explored in other fields such as emotion classification, melanoma classification, and breast mass classification, among others. In this regard, Yoo et al. [40] proposed a joint learning approach using a mini-batch strategy and adaptive fade learning to use the generated images in the classifier with application in visually similar data. Likewise, Zhang et al. [41] and Mabu et al. [42] proposed the use of cycle consistency for classification in retinal pathologies identification and opacity classification in CT scans respectively.

C. Semi-Supervised Image Classification

A common characteristic of medical image datasets is the lack of large annotated sets [43]. During the last few years semi-supervised learning methods have progressed as a good alternative to leverage this large amount of unlabeled data. One of the most common paradigms of semi-supervised learning is the use of Teacher-Student Networks (TSN) [44]. In this type of model, a teacher network is trained on the labeled data, and a student network is trained on the unlabeled data using the predictions given by the teacher. Training in semi-supervised mode allows the student model to learn features from unlabeled datasets [45].

In the endoscopic scenario, few studies have been carried out using semi-supervised learning. Du et al. [46] implemented a semi-supervised contrastive learning method for Esophageal Disease Classification in a small dataset. Golhar et al. [47] proposed the use an unsupervised jigsaw learning method for GI lesion classification obtaining an improvement in accuracy of 9.8% with respect to supervised methods. Guo et al. [48] proposed the use of a combination of a discriminative angular loss and Jensen-Shannon divergence loss for semi-supervised learning for wireless-capsule endoscopic image classification. Shi et al. [49] implemented a TSN network for the 3D reconstruction of stereo endoscopic images.

Recently, semi-supervised GAN-based models have been proposed for image classification in different fields such as natural images and hyper-spectral image classification [50], [51], [52], [53]. However, in the field of endoscopic images it remains an unexplored topic.

Unlike the studies presented in [54], [55], [56], [57], [58] where cycle-consistency translation has been implemented as a way of augmenting their datasets, we use image-translation inside a semi-supervised training loop to improve the classification performance of the unlabeled domain. Furthermore, the methods in which GAN-based semi-supervised methods have been proposed are mainly focused on the classification of images of the same domain.

In this work, we propose a synergic semi-supervised GAN-based method that enables not only the exploitation of unlabeled data but also performs image translation to alleviate the dataset's domain imbalance. This allows the proposed network achieves a better generalization even in an image domain where labels are not available.

III. Methods

Our overall goal is to improve tissue classification of endoscopic bladder images when labels are limited to only one domain, and there is no identical equivalent for every image on each domain. In our case, the endoscopic images are available on WLI and NBI domains, and the labels correspond only to the ones on WLI.

A. Problem Statement

The proposed method consists of three main components; 1) A cycle-consistency translation network to translate every image in the dataset and have equivalent paired images in both domains (NBI and WLI); 2) A teacher network trained on the labeled WLI data; and 3) A multi-input multi-domain classifier trained as student network in a TSN semi-supervised way. A schematic of the proposed model is depicted in Fig. 2.

Let us define a dataset $\mathcal{X} = \mathcal{X}_A \cup \mathcal{X}_B$ composed by the union of two subsets: $\mathcal{X}_A = \{(\boldsymbol{x}_{A1}, \boldsymbol{y}_{A1}), \ldots, (\boldsymbol{x}_{An}, \boldsymbol{y}_{An})\}$ composed by n labeled images $\boldsymbol{x}_i$ belonging to domain $\mathbb{A}$, and $\mathcal{X}_B = \{\boldsymbol{x}_{B1}, \boldsymbol{x}_{B2}, \ldots, \boldsymbol{x}_{Bm}\}$ composed by m unlabeled images $\boldsymbol{x}_j$ belonging to domain $\mathbb{B}$. Initially, a classifier $\mathcal{C}$ is trained in a fully supervised fashion on $\mathcal{X}_A$. This classifier will work as a teacher model $\mathcal{C}_T$ at a later stage. We propose the use of cycle-consistency image translation to deal with the issue of an unpaired and imbalanced dataset. For each image in domain $\boldsymbol{x}_A \in \mathbb{A}$ we will generate an equivalent translation $\hat{\boldsymbol{x}}_{AB} \in \mathbb{B}$, and for every $\boldsymbol{x}_B \in \mathbb{B}$ we will generate an equivalent translation $\hat{\boldsymbol{x}}_{BA} \in \mathbb{A}$. The translated images $\hat{\boldsymbol{x}}_{AB}$ and $\hat{\boldsymbol{x}}_{BA}$ are produced by the generators $\mathcal{G}_{AB}$ and $\mathcal{G}_{BA}$ respectively. An advantage of using cycle-consistency GANs is that an additional image $\hat{\hat{\boldsymbol{x}}}$ is generated, which corresponds to the reconstruction back to the original image. This can be used as additional data to train the student classifier. Therefore for every image $\boldsymbol{x}_A$ we have two extra images $\hat{\boldsymbol{x}}_{AB}$ and $\hat{\hat{\boldsymbol{x}}}_{ABA}$ and the same for $\boldsymbol{x}_B$ where we have $\hat{\boldsymbol{x}}_{BA}$ and $\hat{\hat{\boldsymbol{x}}}_{BAB}$. Then we train a multi-input classifier $\mathcal{C}_S$ which takes as input $\mathcal{C}_S(\boldsymbol{x}_A, \hat{\boldsymbol{x}}_{AB}, \hat{\hat{\boldsymbol{x}}}_{ABA})$ or $\mathcal{C}_S(\boldsymbol{x}_B, \hat{\hat{\boldsymbol{x}}}_{BA}, \hat{\boldsymbol{x}}_{BAB})$, depending on the domain of the input data.

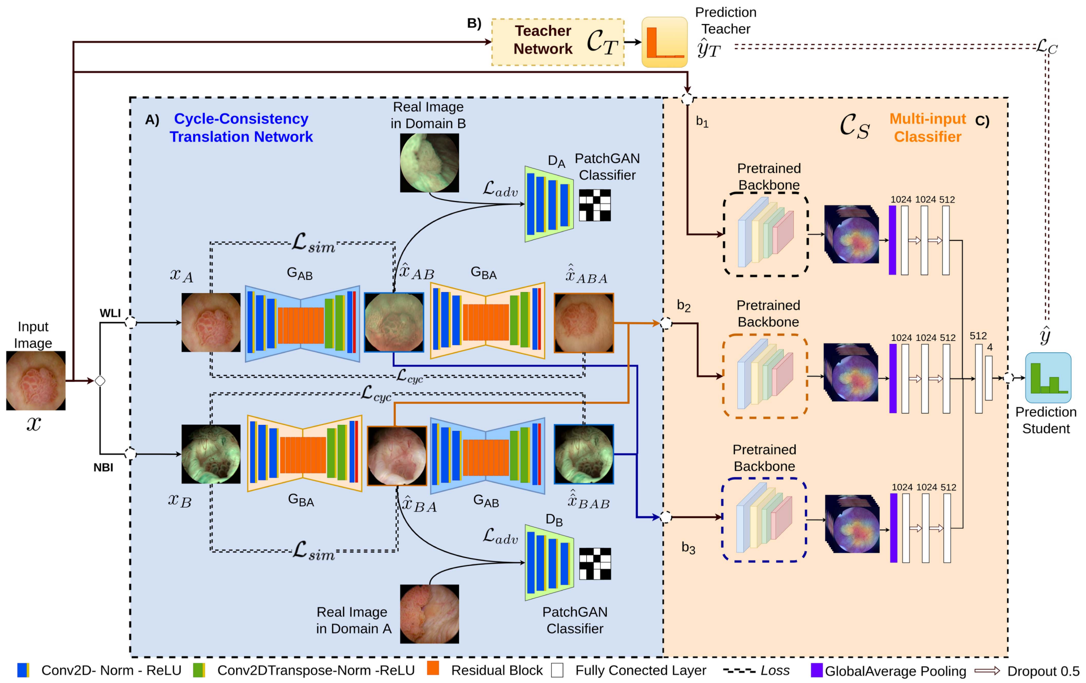


Fig. 2. Proposed method. The network has two main elements. A). Cycle-Consistency Translation Network that translates the image from NBI to WLI and vice-versa. B). Teacher network. C). Multi-input network that performs the tissue classification task based on the features from both image modalities. The classification makes use of backbone networks that extract the features from each of the inputs to the classifier. The features are processed using Fully Connected (FC) layers which later are concatenated to perform the prediction in the final layer.

B. Cycle-Consistency Translation Network

The unpaired image-to-image translation network is a generative adversarial network based on the *Cycle*GAN architecture [59]. Two generators $\mathcal{G}_{AB}$ and $\mathcal{G}_{BA}$ are trained to learn the mappings between the domains $\mathbb{A}$ = WLI and $\mathbb{B}$ = NBI, such that $\mathcal{G}_{AB} : \mathbb{A} \rightarrow \mathbb{B}$ and $\mathcal{G}_{BA} : \mathbb{B} \rightarrow \mathbb{A}$. $\mathcal{D}_A$ and $\mathcal{D}_B$ are the two discriminators trained two distinguish between the real and fake images of each domain. The proposed model uses three main losses, the adversarial loss $\mathcal{L}_{adv}$, the cycle consistency loss $\mathcal{L}_{cyc}$ and a similarity loss $\mathcal{L}_{sim}$.

The cycle loss $\mathcal{L}_{cyc}$ is defined as

$$\mathcal{L}_{cyc}(\mathcal{G}_{pq}, \mathcal{G}_{pq}, \boldsymbol{x}_p) = \mathbb{E}_{x_p}||\boldsymbol{x}_p - \mathcal{G}_{qp}(\mathcal{G}_{pq}(\boldsymbol{x}_p))|| \tag{1}$$

where the indexes p, q represent the domain of the image and the domain to which is translated. The adversarial loss for each generator $\mathcal{G}_{pq}$ and discriminator $\mathcal{D}_p$ is defined as

$$\begin{aligned}\mathcal{L}_{adv}(\mathcal{G}_{pq}, \mathcal{D}_p) &= \mathbb{E}_{\hat{x}_p}[log(\mathcal{D}_p(\hat{\boldsymbol{x}}_p))] \\ &\quad + \mathbb{E}_{x_p}[log(1 - D_p(\mathcal{G}_q(\boldsymbol{x}_p)))]\end{aligned} \tag{2}$$

To preserve the fine-grain details, such as the capillaries and inner blood vessels, that are related to the intrinsic pathology of each image domain and which are an essential visual cue for diagnosis assessment, we propose the addition to the cycle-consistency network a similarity loss $\mathcal{L}_{sim}$. This is defined as:

$$\begin{aligned}\mathcal{L}_{sim}(\mathcal{G}_{AB}, \mathcal{G}_{BA}) &= \left[1 - \sum_{i}^{N} F\left(\hat{\boldsymbol{x}}_{Ai}, \mathcal{G}_{AB}(\boldsymbol{x}_{Ai})\right)\right] \\ &\quad + \left[1 - \sum_{i}^{N} F\left(\hat{\boldsymbol{x}}_{Bi}, \mathcal{G}_{BA}(\boldsymbol{x}_{Bi})\right)\right]\end{aligned} \tag{3}$$

where $\boldsymbol{x}_A \in \mathbb{A}$ and $\boldsymbol{x}_B \in \mathbb{B}$ correspond to the images form the $\mathbb{A}$ and $\mathbb{B}$ domains and the ith refers index over the a set of images of N elements. $\hat{\boldsymbol{x}}_A$ and $\hat{\boldsymbol{x}}_B$ correspond to the translated images by the generators. $F(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is the structural similarity (SSIM) between images x and $\hat{x}$ proposed in [60] as:

$$F(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)} \tag{4}$$

Where $\sigma_{x,\hat{x}}$ is the covariance between x and $\hat{x}$:

$$\sigma_{x,\hat{x}} = \frac{1}{m-1}\sum(x_j - \mu_x)(\hat{x}_j - \mu_{\hat{x}}) \tag{5}$$

m is the number of pixels; x_j and $\hat{x}_j$ are the jth pixel of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ respectively; μ_x, $\mu_{\hat{x}}$ and σ_x and $\sigma_{\hat{x}}$ are the mean intensities and standard deviations of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, and c_1 and c_2 are stabilizing constants to avoid singularities when $\mu_x^2 + \mu_{\hat{x}}^2 \approx 0$ and $\sigma_x^2 + \sigma_{\hat{x}}^2 \approx 0$ respectively.

The overall objective function of the generative network is then defined as

$$\begin{aligned}\mathcal{L}(\mathcal{G}_{AB}, \mathcal{G}_{BA}, \mathcal{D}_A, \mathcal{D}_B) &= \mathcal{L}_{adv}(\mathcal{G}_{AB}, \mathcal{D}_A) \\ &\quad + \mathcal{L}_{adv}(\mathcal{G}_{BA}, \mathcal{D}_B) + \lambda_1\mathcal{L}_{sim}(\mathcal{G}_{AB}, \mathcal{G}_{BA}) \\ &\quad + \lambda_2\mathcal{L}_{sim}(\mathcal{G}_{AB}, \mathcal{G}_{BA}) + \lambda_3\mathcal{L}_{cyc}(\mathcal{G}_{AB}, \mathcal{G}_{BA}, x_A) \\ &\quad + \lambda_4\mathcal{L}_{cyc}(\mathcal{G}_{BA}, \mathcal{G}_{AB}, x_B)\end{aligned} \tag{6}$$

where λ_i are the hyper-parameters that balance the impact of the losses. The generators are trained to minimize the overall function and the discriminators to maximize it. The proposed *Cycle*GAN with Similarity loss is termed CSi-GAN in the remainder of this paper, and the case in which $\lambda_1 = \lambda_2 = 0$ it reverts to the classical *Cycle*GAN.

C. Semi Supervised Classification

Initially, the teacher model C_T is trained on WLI images in a fully supervised way. This could be seen as disconnecting the branch that goes from the input image $\boldsymbol{x}$ to the Cycle-Consistency Translation Network in Fig. 2, and training the network to optimize (7) substituting the $\hat{\boldsymbol{y}}_{Ti}$ pseudo-labels with the labels $\boldsymbol{y}_{Ai}$ from set $\mathcal{X}_A$. Afterward, the student model C_S is trained using the labeled and unlabeled data using the predictions $\hat{\boldsymbol{y}}_T$ obtained from the teacher. The student network corresponds to a multi-input classifier that takes 3 images as input $C_S(\boldsymbol{x}, \hat{\boldsymbol{x}}, \hat{\hat{\boldsymbol{x}}})$ as depicted in Fig. 2-(C). The first one $\boldsymbol{x}$ is the original image from either WLI ($\boldsymbol{x}_A$) or NBI ($\boldsymbol{x}_B$) domains, the other two images correspond to the ones generated by the generators $\mathcal{G}_{AB}$ and $\mathcal{G}_{BA}$ respectively. In the case of the branch that takes as input $\boldsymbol{x}$, random data augmentation operations are applied which include random crop, random rotation, and flipping. Backbone networks b_1, b_2, and b_3, are used to extract the features of each of the 3 input images. In our case, we used as backbone ResNet-101 trained on *ImageNet*. The extracted features from each of the backbones are processed separately using a shallow network composed of 3 Fully Connected (FC) layers. The outputs from these layers are concatenated together, from which finally the class prediction is performed in the final layer. The classifier was trained to optimize the categorical cross-entropy loss defined as:

$$\mathcal{L}_C(\hat{\boldsymbol{y}}_{Ti}, \hat{\boldsymbol{y}}_{\boldsymbol{i}}) = -\sum_i \hat{\boldsymbol{y}}_{Ti} \cdot log(\hat{\boldsymbol{y}}_{\boldsymbol{i}}) \tag{7}$$

where $\hat{\boldsymbol{y}}_i$ is the predicted output from the student model, $\hat{\boldsymbol{y}}_{Ti}$ is the corresponding pseudo-label provided by the teacher network, and i refers to the index over the classes.

D. Dataset

For this study, endoscopic videos from 23 patients undergoing TURBT were collected, as well as the respective histopathological analysis from the resected lesions. The matching between the visual data and the histological results was done with the aid of an expert surgeon. The matching was performed by analyzing frame-by-frame the videos. The sections of the bladder from which lesions were resected during the surgical intervention were then identified. To avoid ambiguities of having multiple lesions of multiple types, only the frames in which individual lesions appeared were used in the dataset. This procedure was performed on all the WLI video clips as well as 3 patients with NBI video data. In total 4 classes were defined. Taking into consideration the general classification of BC as defined in [2] by the WHO and the International Society of Urological Pathology (ISUP), two categories were considered for cancerous tissue: Low-Grade Cancer (LGC) and High-Grade Cancer (HGC). Additionally, 2 extra categories were considered for No Tumor Lesion (NTL) which comprehends cystitis, caused by infections or other inflammatory agents, and Non-Suspicious Tissue (NST). The detailed statistics of the dataset are shown in Table I.

The videos were acquired at the European Institute of Oncology (IEO) at Milan, Italy. Each patient signed an informed consent document approved by the IEO and in accordance with the Helsinki Declaration. No personal data was recorded.

To determine if the use of more data helps to achieve better generalization when training the GAN networks, we used additional data from the datasets presented in [14], [27] which contains endoscopic images from colonoscopy in NBI and WLI domains, and [61] which contains unlabeled data from TURBT as well in NBI and WLI domains.

TABLE I
Composition of the Dataset Considering Two Light Modalities; White Light Imaging (WLI) and Narrow Band Imaging (NBI)

Tissue type	No. of patient cases	No. of images		
		WLI	NBI	Total
HGC	8	386	64	469
LGC	9	454	145	647
NST	5	439	75	504
NTL	5	97	37	134
Total	23*	1433	321	1754

*The total number of patient cases does not correspond to the sum of the second column since some of the patients had more than one type of lesion.

TABLE II
Dataset Composition Used for Training the GAN Models

Dataset type	composition	No. of images		
		NBI	WLI	Total
$\mathbb{D}_1$	$\mathbb{D}_A$	1036	228	1264
$\mathbb{D}_2$	$\mathbb{D}_B \cup \mathbb{D}_C \cup \mathbb{D}_D$	4592	2512	7104
$\mathbb{D}_3$	$\mathbb{D}_A \cup \mathbb{D}_2$	5628	2740	8368

$\mathbb{D}_A$: our dataset. $\mathbb{D}_B$: dataset described on [62].
$\mathbb{D}_C$: dataset described on [28]. $\mathbb{D}_D$: dataset described on [14]

$\mathbb{D}_1$ corresponds to our dataset described in section. III-D. $\mathbb{D}_2$ corresponds to a dataset composed only of external sources. $\mathbb{D}_3$ corresponds to the union of all the previously mentioned datasets.

E. Model Implementation

The model was trained in three steps. First, the cycle consistency GAN was trained for 150 epochs with an initial learning rate of 2e$-$4 and batch size of 1. The λ hyperparameters were set to λ_1=λ_2=2.0, and λ_3=λ_4=1.0 The second step consisted of training the teacher classifier using the labeled dataset $\mathcal{X}_A$. Once the GAN model and the teacher networks were trained, the multi-input classifier was trained setting the initial learning rate at 1e$-$5 using a batch size of 32. The models were implemented using Tensorflow 2.5 in Python 3.6 and deployed on an Nvidia GeForce GTX 1080 GPU. The training of the classifiers was repeated 10 times for each of the different experiments carried out in this study.

For performance benchmarking of the classifiers, a hold-out strategy was used, 4 patient cases randomly chosen were held as test dataset. The rest of the dataset was divided randomly in a 75/25 ratio for training/validation. In the case of the GAN models, only the train dataset used for supervised classification was used during the training of the different combinations described in Table II. For the semi-supervised training apart from using the labeled WLI images and unlabeled NBI, all NBI cystoscopy images described in [61] were added to the training dataset. The test dataset for the semi-supervised task remained the same as the one used to test the performance of the teacher model.

F. Evaluation Protocol

Each of the different modules that comprise the proposed method was evaluated separately, and the best components of each one were chosen.

In contrast with other DL models that are trained to minimize a loss function, GAN models are trained to converge to an equilibrium between the generator and the discriminator networks. For this reason, there is no objective loss function to train this type of model, and compare their performance objectively [50]. However, there are some quantitative techniques that have been proposed to assess the performance of GAN models [62].

1) Quantitative Evaluation of the Generators: Generator models are usually evaluated based on the quality of the images they generate. However, this type of evaluation might not fully show the performance of the models and might be subjective due to biases of the reviewer [62]. In this regard, some authors have proposed the use of different metrics such as the Inception score, to quantitatively evaluate the quality of the generated images [50]. In our specific case, we have the limitation that the dataset does not correspond to natural images, such as the ones on *ImageNet* dataset, and therefore we can not apply the Inception score directly. We use instead the FrÃ©chet Inception Distance (FID) proposed in [63], to quantify the performance of each generator trained and defined as:

$$d^2\left((\mathbf{m},\mathbf{C}),(\mathbf{m}_\omega,\mathbf{C}_\omega)\right) = \|\mathbf{m}-\mathbf{m}_\omega\|_2^2 + Tr\left(\mathbf{C}+\mathbf{C}_\omega - 2(\mathbf{C}\mathbf{C}_\omega)^{1/2}\right) \tag{8}$$

were $\mathbf{m}$, $\mathbf{C}$ are the mean and covariance obtained from the last pooling layer of an Inception model using sample images produced by the generative model respectively, and $\mathbf{m}_\omega$, $\mathbf{C}_\omega$ are the corresponding ones using images from the original dataset.

We also analyze how the amount of data affects the quality of the images and the classifiers' performance. For this purpose, we use 3 different combinations of datasets coming from 4 different sources. The datasets composition is shown in Table II.

To measure the sensitivity of the models depending on the amount of data used, we analyze the sensitivity to noise for each of the generative models trained on the different datasets as proposed in [64]. We added zero-mean Gaussian noise $N(0,\sigma)$ in a range of $\sigma = [0.025, 0.05, 0.075, 0.1, 0.2]$ to the translation result before reconstruction. We compute the Mean Square pixel Error (MSE) of the reconstructed image with respect to the original image $\boldsymbol{x}_i$ and calculate the sensitivity (SN) using the equation:

$$SN = \frac{1}{N}\sum_{i=1}^{N} MSE(\mathcal{G}_p(\mathcal{G}_q(x_i) + N(0,\sigma) - \boldsymbol{x}_i) \tag{9}$$

We compared the sensitivity for each of the generators in the proposed Cycle Similarity network (CSi-GAN) and the baseline *CycleGAN*.

2) Evaluation by Medical Specialists: Once the different GAN models were trained, the one with the best FID score was selected as the one to be used for human evaluation. With this analysis, we intended to confirm that the quality of the generated images is good enough to deceive experts, as well as to have a baseline to compare the classification performance of the models with respect to the ones from specialists.

To qualitatively evaluate the utility of the images an online survey was set up where medical experts were asked to complete two tasks. In the first task, 20 pairs of randomly selected images were shown to the participants. Each image pair corresponded to two images from the same domain; one of them was an original image taken with the endoscope while the other corresponded to a translated image by the GAN. The participants were asked to identify which one was the original one, and which one was the generated one. For this task, NBI and WLI image pairs were evenly distributed with 10 samples for each case. In the second task, 40 pairs of images were shown to the participants. The clinicians were asked to classify the images according to the 4 classes explained in Section III-D. Each image pair corresponded to one of the following options distributed in a 50/50 ratio: 1) A pair of images that showed the same anatomical region at different times. In this case, the pair of images could correspond to two images of the same region and the same domain or two images of the same region but with a different domain, i.e. (NBI, NBI), (WLI, WLI) and (NBI, WLI). Each of the possible cases was evenly distributed. 2) In the second option, again two images were shown that correspond to the same anatomical region at different times. However, in this case one of the images was domain translated. The images used in this task were randomly chosen, taking into consideration having an even distribution of the 4 different tissue classes. Image pairs from options 1) and 2) were randomly ordered across the survey.

3) Evaluation of the Classifiers: Once the GAN models were trained, we incorporate them into the general workflow using them as the base backbone to produce the multi-domain input images to feed the student classifier. The training was performed first in a fully supervised manner and then in a semi-supervised way using the previously trained teacher. To select the teacher model, diverse pre-trained models previously used in the literature were trained and the one with the best performance metrics was chosen as the teacher. We also performed ablation studies as well to demonstrate the utility of each of the elements of the proposed method. In the final stage, we train the multi-input classifier in a fully supervised way, using each of the previously trained generative models to determine whether there is a correlation between the classification performance and the quality of the generated images.

G. Evaluation Metrics for Classification

To evaluate the classification performance of the proposed method we used the metrics: accuracy (Acc), precision ($Prec$), recall (Rec), and F1-score. Additionally, as proposed in [48] we also evaluated the model using Matthews correlation coefficient (MCC) and Cohen's kappa (CK) statistic which has shown to be effective to benchmark diagnosis reliability of classifiers [65]. Mann Whitney U-test was used to determine the statistical significance. In the case of the user's experiments, the same metrics were used to evaluate their performance. Additionally, for the users' task of identifying the real images from the fake ones, the Area Under the Curve of the Receiver Operating Characteristic curve AUC was used.

IV. Results and Discussion

This section is divided into two main subsections. First, we evaluate the performance of the image-translation network quantitative and qualitatively. Then we proceed to analyze the results of the classification network and the influence that the quality of the generated images has on the overall system, as well as the different components of the system.

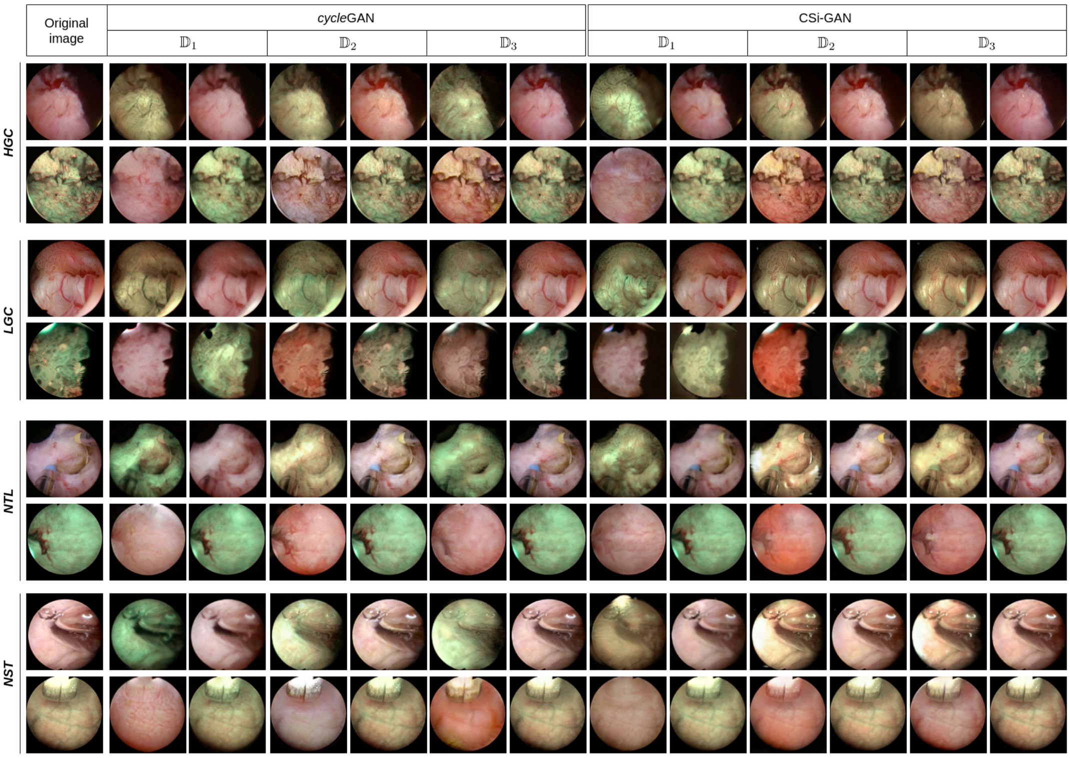


Fig. 3. Samples of the generated images for the 4 classes on the 2 domains using each of the GAN models. For each model trained on the 3 different datasets ($\mathbb{D}_1$, $\mathbb{D}_2$, $\mathbb{D}_3$) two images are shown: 1) the translated image to the complementary domain, and 2) the reversed translation back to its original domain.

A. Evaluation of the GAN Models

The first set of results corresponds to the qualitative assessment of the synthetically generated images. Samples of randomly chosen generated images by the different GAN models trained are shown in Fig. 3. A visual comparison shows that the amount and diversity of training data improve the quality of the images. We can observe that the addition of data helps the network learn the existence of other objects which do not correspond to the anatomical structures in the body, such as tools or bubbles. This shortcoming where the networks tend to disappear external structures by coloring them with the same hue as the rest of the background is more perceptible when models are trained with small datasets ($\mathbb{D}_1$). Furthermore, in these cases, the network also presents some noticeable flaws since sometimes the generated images present black dots scattered at diverse points. Nevertheless, the use of only external data ($\mathbb{D}_2$) also alters the hue of the translation. This could be linked to the fact that the external data comes mainly from GI images which present different tints and anatomical formations than the ones present in the bladder. In general, for both cases *cycle*GAN and CSi-GAN the use of the more general dataset ($\mathbb{D}_3$) which comprises data from the same anatomical target and external data produce the best quality images. However, still some image artifacts such as specularities, reflections, interlacing, etc. appear in the generated images without being present in the original one. The most significant improvement comes from using the $\mathcal{L}_{sim}$ loss to train the GANs. The fine-grain details, such as small vessels, are better preserved and highlighted after the translations, and it also helps to reduce the amount of noise in the image. Similar behaviors can be observed in the video material attached to this manuscript.

TABLE III
FID Scores and AUC of the Sensitivity Curves for Each of the GAN Models Trained on the Different Datasets

model	dataset	FID		AUC	
		$\mathcal{G}_{AB}$	$\mathcal{G}_{BA}$	$\mathcal{G}_{AB}$	$\mathcal{G}_{BA}$
*Cycle*GAN	-$\mathbb{D}_1$	146.74	214.93	295.303	176.99
	-$\mathbb{D}_2$	130.09	169.72	82.55	91.13
	-$\mathbb{D}_3$	138.79	164.24	113.69	119.57
CSiGAN	-$\mathbb{D}_1$	73.96	117.65	245.95	125.19
	-$\mathbb{D}_2$	54.33	72.13	81.76	80.18
	-$\mathbb{D}_3$	**35.73**	**37.67**	**78.87**	**52.32**

The results are divided in terms of the two generators G_{AB} and G_{BA}. The numbers in bold indicate the cases that obtained the best metrics.

1) Quantitative Evaluation of the GAN: To evaluate the quality of the images generated by the GAN models the FID score and the AUC of the sensitivity curve were used. The results obtained for both metrics are shown in Table III. The model that obtains the best metrics for both cases, i.e. lower values, is the proposed CSi-GAN when trained on $\mathbb{D}_3$. In the case of FID score there is a clear difference between *Cycle*GAN and CSi-GAN regardless of the dataset used for training, with CSi-GAN obtaining in general better results. In the case of the AUC of the Sensitivity curve, the difference between the two models is not that obvious. This could be associated with the fact that neither of the networks is designed from the origin to be noise-resistant. However, there is a clear tendency that the addition of data makes CSiGAN more resistant to the addition of noise than its counterpart *Cycle*GAN. This might be related to

TABLE IV
AVERAGE RESULTS ± STANDARD DEVIATION FROM THE SPECIALIST EVALUATION REGARDING THEIR ABILITY TO DISCERN BETWEEN REAL AND GENERATED IMAGES

group (n)	Translation type	Acc	$Prec$	Rec	AUC
ES (15)	WLI → NBI	0.66±0.13	0.66±0.18	0.8±0.20	0.59±0.14
	NBI → WLI	0.50±0.14	0.44±0.15	0.75±0.19	0.55±0.13
	ALL	0.57±0.09	0.57±0.10	0.66±0.12	0.59±0.09
RE (5)	WLI → NBI	0.66±0.00	0.83±0.16	0.60±0.20	0.67±0.02
	NBI → WLI	0.40±0.10	0.34±0.050	0.50±0.00	0.41±0.08
	ALL	0.52±0.05	0.51±0.05	0.55±0.11	0.52±0.44

Results are divided in terms of the two different groups: expert surgeon (ES)and resident (RE), and by the type of translation performed by each generator network i.e. WLI → NBI, NBI → WLI as well as the overall performance (all) of the gan.

the fact that even if the addition of more data helps *Cycle*GAN to generalize better in domain translation the lack of a structural loss inhibits it to discern properly between the correct information to produce a satisfactory translation, and the information that seems useful but is just noise. This could also explain the reason why *Cycle*GAN obtains better metrics when trained on dataset $\mathbb{D}_2$ than on $\mathbb{D}_3$ since the quality of the images of $\mathbb{D}_2$ is higher and less noisy.

2) Evaluation by Medical Specialists: In order to perform a more exhaustive analysis, a protocol was implemented to acquire feedback from expert clinicians in the field of endoscopy as described in Section III-F2. A total of 20 physicians from 10 different institutions familiar with TURBT participated in the study. Of this, 15 corresponded to Expert Surgeons (ES) and 5 to Residents (RE). For this analysis we choose the generative model which obtained the best FID score and AUC values, i.e. CSi-GAN trained on dataset $\mathbb{D}_3$, to generate the synthetic images.

The results regarding the ability of surgeons to discern between real and synthetic images are shown in Table IV. The results are split in 3 categories to evaluate separately each translation (WLI → NBI and NBI → WLI) and therefore each generator independently, as well as the overall performance of the GAN (ALL). For both groups of participants (ES and RE), the results show slightly better results in the translation WLI → NBI for all metrics. This might be related to the fact that there are more sample images in the WLI training dataset than in the NBI and therefore the generator $\mathcal{G}_{AB}$ is able to generalize better and produce better quality images than its counterpart $\mathcal{G}_{BA}$. The overall AUC for ES is 0.59 and 0.52 for RE, meaning that their performance is marginally better than what a random binary classifier could achieve, confirming that the quality of the generated images is good enough to trick experts in the area.

Concerning the tissue classification task, results are shown in Fig. 4. In the case of Acc there was an average improvement of 8% when using a pair of a real image and a synthetic one than when only 2 real images were shown. In the case of $Prec$ the improvement was 19%, while no improvement or decrease was observed in the case of Rec. For the F-1 score and MCC the improvements were 16% and 17% respectively. However, no statistical significance was found. This goes in accordance with the results obtained in the previous analysis, meaning that the generated images do not affect the specialist's performance on tissue classification.

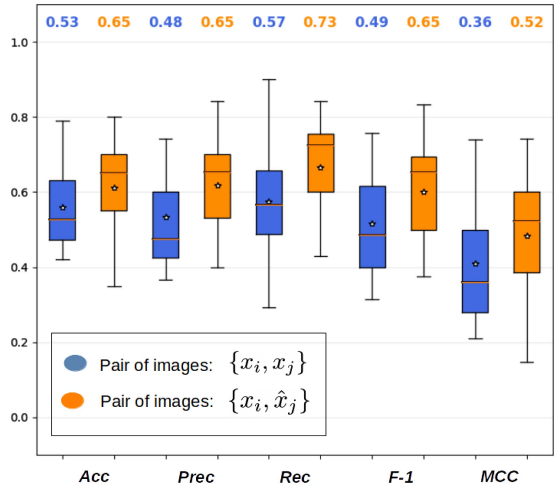


Fig. 4. Box plot comparison of the surgeons performance in the tissue classification task. Blue boxes correspond to the case in which surgeons were shown a pair of real images $\{x_i, x_j\}$. Orange boxes correspond to cases in which a pair consisting of a real image x and its translation $\hat{x}$ to the opposite domain $\{x_i, \hat{x}_j\}$, are shown.

B. Tissue Classification Evaluation

Results regarding tissue classification are divided into three parts. First, we show that the use of our proposed GAN method for image translation improves in general the performance of tissue classification using different backbones previously used in the literature as simple fine-tuned classification networks. Next, we show that the use of semi-supervised learning, in general, improves further the classification performance. Finally, we perform an ablation analysis of the proposed model.

1) GAN-Based Tissue Classification: To test the generalization of our method, we compare the use of different networks (VGG16, VGG19, Inception V3, Desenet, ResNet-50, and ResNet-101) trained in a fine-tuning fashion against the implementation of these same networks in our GAN-based classification method. CSi-GAN trained on $\mathbb{D}_3$ was chosen as the as the translation network. Results in terms of ACC, MCC and F-1 score are shown in Table V. Overall the use of the proposed GAN-based method obtains better metrics than the baseline networks. In the majority of the cases, there is little improvement or no improvement when the input image is in the WLI domain. This uneven behavior in terms of the classification improvement might be related to the fact that WLI images are more similar to the natural images dataset in which the models were originally pre-trained (*ImageNet*). However, there is a noticeable improvement when it comes to the classification of NBI images where most of the base-line shows poor performances.

2) Semi-Supervised Classification: We compared the use of GAN-based classification trained in a fully supervised way against the use of semi-supervised classification. In both cases, only the Multi-Input classifier weights were trained while the ones of the Cycle-Consistency Network remained constant. For these experiments, CSi-GAN pre-trained on each of the $\mathbb{D}_k$ datasets were used. The results of these experiments are shown in Fig. 5 in terms of ACC, F-1 score, and MCC. On average the improvement, in terms of ACC, F-1 score, and MCC, of using CSiGAN trained in a fully supervised way against the training in

TABLE V
COMPARISON OF USING DIFFERENT PRE-TRAINED MODELS IN THE PROPOSED GAN-BASED MULTI-INPUT CLASSIFIER

model	test data	*ACC*			F-1			*MCC*		
		baseline	GAN-based	p-val	baseline	GAN-based	p-val	baseline	GAN-based	p-val
VGG19	NBI	0.667±0.030	0.821±0.058	0.007	0.245±0.057	0.529±0.167	0.003	0.272±0.119	0.673±0.059	0.003
	WLI	0.653±0.048	0.667±0.033	0.789	0.675±0.034	0.716±0.057	0.060	0.487±0.054	0.564±0.084	0.298
	ALL	0.661±0.033	0.684±0.031	0.286	0.649±0.022	0.649±0.052	0.797	0.567±0.038	0.572±0.053	0.298
VGG16	NBI	0.692±0.056	0.744±0.075	0.018	0.409±0.144	0.409±0.174	0.325	0.010±0.212	0.376±0.237	0.060
	WLI	0.720±0.022	0.740±0.025	0.014	0.641±0.046	0.716±0.025	0.002	0.610±0.030	0.632±0.035	0.006
	ALL	0.714±0.017	0.741±0.028	0.002	0.648±0.046	0.741±0.023	0.001	0.602±0.024	0.634±0.036	0.001
Inception V3	NBI	0.833±0.028	0.833±0.044	0.891	0.530±0.151	0.685±0.063	0.011	0.591±0.112	0.602±0.065	0.893
	WLI	0.713±0.031	0.733±0.017	0.325	0.645±0.031	0.676±0.028	0.016	0.624±0.038	0.636±0.018	0.408
	ALL	0.743±0.026	0.751±0.011	0.280	0.643±0.028	0.68±0.025	0.002	0.658±0.014	0.661±0.033	0.633
Densenet	NBI	0.641±0.041	0.718±0.086	0.054	0.240±0.054	0.295±0.181	0.048	0.279±0.095	0.407±0.129	0.033
	WLI	0.763±0.036	0.767±0.049	0.879	0.782±0.049	0.743±0.054	0.761	0.679±0.070	0.725±0.055	0.879
	ALL	0.767±0.031	0.772±0.039	0.675	0.759±0.04	0.780±0.037	0.447	0.684±0.054	0.692±0.051	0.820
ResNet-50	NBI	0.718±0.038	0.809±0.053	0.002	0.316±0.058	0.633±0.176	0.001	0.390±0.185	0.642±0.152	0.004
	WLI	0.830±0.010	0.860±0.014	0.003	0.806±0.037	0.820±0.018	0.307	0.769±0.028	0.788±0.057	0.391
	ALL	0.811±0.017	0.857±0.017	0.001	0.826±0.014	0.842±0.016	0.008	0.783±0.020	0.804±0.021	0.006
ResNet-101	NBI	0.744±0.085	**0.862±0.046**	0.011	0.452±0.242	**0.713±0.174**	0.016	0.547±0.196	**0.757±0.081**	0.008
	WLI	0.861±0.027	**0.867±0.025**	0.327	0.804±0.028	**0.832±0.029**	0.595	0.801±0.036	**0.806±0.031**	0.304
	ALL	0.831±0.031	**0.865±0.026**	0.038	0.831±0.062	**0.854±0.029**	0.114	0.766±0.040	**0.816±0.026**	0.030

The average ± standard deviation for each metric is presented in terms of the type of data in the test dataset (WLI and NBI) and the combination of both (all), for each of the models. The numbers in bold indicate the cases that obtained the best metrics.

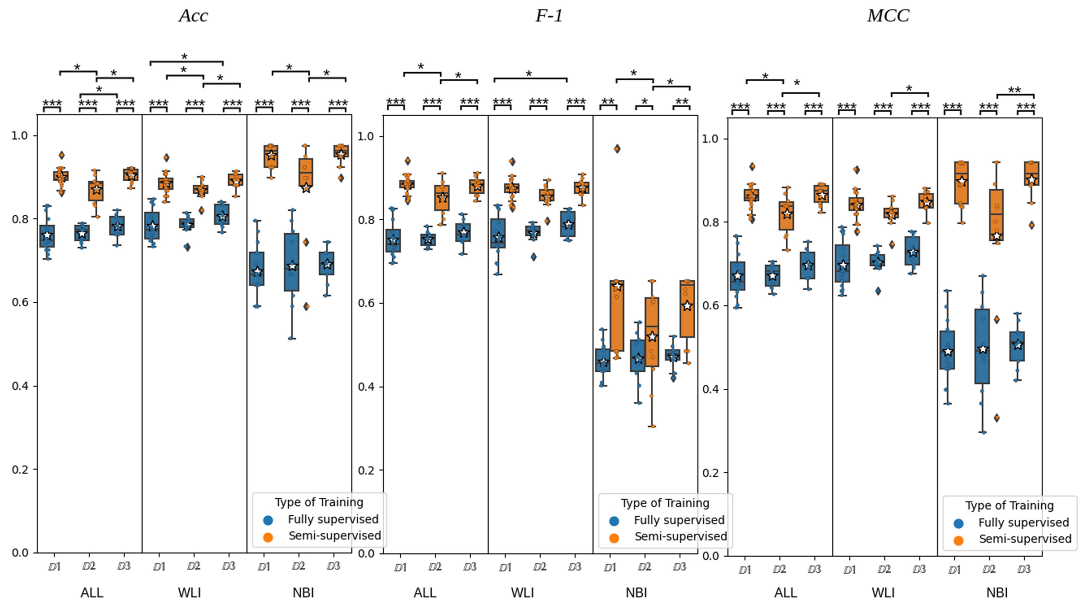
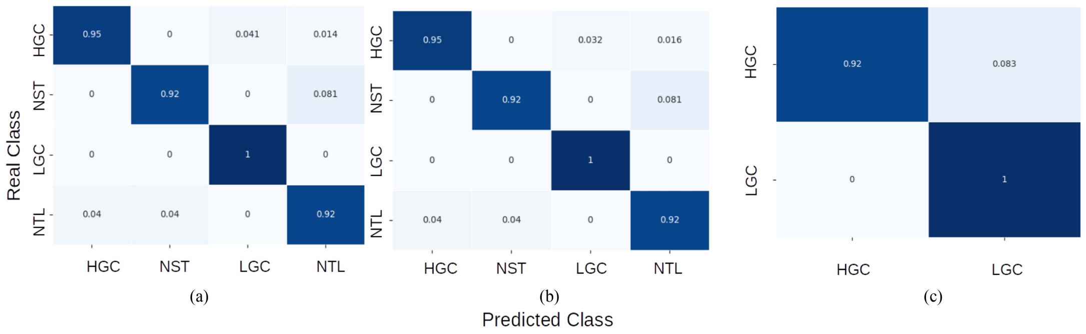


Fig. 5. Boxplots comparison of Acc, $F-1$ score and MCC of the proposed model trained in fully supervised vs semi-supervised way using CSi-GAN pre-trained on $\mathbb{D}_1$, $\mathbb{D}_2$ and $\mathbb{D}_3$. The results for each metric are divided in terms of the type of data in the test dataset (WLI and NBI) and the combination of both of them (ALL). The statistical significance using Mann Whitney U-test is denoted with $* : p < 0.05$, $** : p < 0.01$, $*** : p < 0.001$.

a semi-supervised fashion was of 8%, 6%, and 9% respectively. This shows the potential of using GAN-based semi-supervised learning for bladder tissue classification. The confusion matrices of the best model obtained are shown in Fig. 6.

3) Ablation Results: In this case, we made a comparison between the base model, the proposed CSiGAN model trained in a fully supervised way, and in a semi-supervised way (Se-CSiGAN). We also analyzed the influence that each of the inputs of the multi-domain classifier model has. For this purpose, we trained the network with each of the individual branches (b_1, b_2, b_3) separately. The statistical significance was calculated with respect to the base-model ResNet-101. Classification results obtained by medical experts, stratified between specialists and residents are shown as a reference point. The results of the ablation experiments are shown in the Tables VI and VII. From these results, we can see that in general, all the models obtain better results than the specialists, and the major improvement comes from the use of a semi-supervised approach. However, the improvement obtained in the domain for which there are no labels when using domain translation is also noticeable. As expected, the integration of both results in the best performance, and improves considerably the detection of classes that are underrepresented. This behavior is more clearly noticeable in the case of the NTL class which in our dataset has the smallest number of samples and in contrast to NST could be easily misclassified as a tumorous lesion.

An additional analysis was performed in order to determine if the quality of the GAN-translated images influence the classifier



(a) (b) (c)

Fig. 6. Confusion matrices of the best model obtained. a) Analysis on the complete test data (WLI + NBI). b) Analysis only on the WLI test data. c) Analysis on the NBI data. Is important to notice that due to the scarcity of annotated NBI data, the NBI test dataset was composed only of HGC and LHC images.

TABLE VI
ABLATION RESULTS

method	UD	DT	test data	Accuracy	p-val	Precision	p-val	Recall	p-val	F-1	p-val	MCC	p-val	CK	p-val
residents	-	-	ALL	0.553±0.116	-	0.521±0.115	-	0.587±0.164	-	0.504±0.134	-	0.405±0.158	-	0.385±0.157	-
specialist	-	-	ALL	0.579±0.111	-	0.542±0.113	-	0.607±0.16	-	0.523±0.132	-	0.424±0.153	-	0.418±0.151	-
baseline (ResNet-101)	✗	✗	ALL	0.831±0.031	-	0.843±0.019	-	0.831±0.062	-	0.831±0.031	-	0.766±0.04	-	0.762±0.044	-
			WLI	0.861±0.027	-	0.868±0.024	-	0.858±0.031	-	0.804±0.028	-	0.801±0.036	-		
			NBI	0.744±0.085	-	0.611±0.210	-	0.85±0.095	-	0.452±0.242	-	0.547±0.196	-		
CSi-GAN-b_2 (FS)	✗	✓	ALL	0.627±0.038	0.003	0.610±0.036	0.001	0.592±0.042	0.001	0.593±0.042	0.001	0.472±0.056	0.001	0.47±0.057	0.001
			WLI	0.610±0.030	0.003	0.587±0.030	0.001	0.572±0.032	0.001	0.565±0.034	0.001	0.455±0.042	0.001		
			NBI	0.692±0.073	1.0	0.549±0.14	0.958	0.806±0.112	0.265	0.529±0.153	0.645	0.441±0.19	0.327		
CSi-GAN-b_3 (FS)	✗	✓	ALL	0.688±0.026	0.001	0.706±0.024	0.001	0.691±0.026	0.001	0.700±0.023	0.001	0.563±0.037	0.001	0.561±0.036	0.001
			WLI	0.700±0.025	0.001	0.723±0.028	0.001	0.723±0.019	0.001	0.705±0.023	0.001	0.610±0.030	0.001		
			NBI	0.641±0.058	0.114	0.487±0.112	0.287	0.840±0.084	0.61	0.404±0.067	0.391	0.483±0.069	0.298		
CSi-GAN (FS)	✗	✓	ALL	0.865±0.020	0.038	0.849±0.017	0.210	0.853±0.0211	0.064	0.854±0.029	0.14	0.816±0.026	0.030	0.812±0.028	0.025
			WLI	0.867±0.025	0.327	0.851±0.025	0.414	0.844±0.029	0.595	0.838±0.029	0.595	0.806±0.032	0.304		
			NBI	0.872±0.046	0.011	0.839±0.023	0.771	0.921±0.054	0.137	0.713±0.174	0.016	0.757±0.081	0.008		
baseline semi-supervised	✓	✗	ALL	0.868±0.019	0.018	0.853±0.024	0.077	0.856±0.02	0.059	0.849±0.021	0.028	0.817±0.026	0.024	0.815±0.026	0.017
			WLI	0.863±0.015	0.731	0.864±0.016	0.926	0.841±0.021	0.239	0.847±0.017	0.476	0.809±0.021	0.598		
			NBI	0.803±0.075	0.027	0.615±0.146	1.0	0.848±0.058	0.082	0.614±0.16	0.456	0.835±0.154	0.072		
SeCSi-GAN	✓	✓	ALL	**0.905±0.026**	0.001	**0.885±0.027**	0.005	**0.892±0.031**	0.004	**0.889±0.031**	0.002	**0.867±0.036**	0.001	**0.866±0.037**	0.001
			WLI	**0.897±0.016**	0.001	**0.887±0.019**	0.012	**0.895±0.022**	0.005	**0.889±0.020**	0.001	**0.856±0.022**	0.002		
			NBI	**0.923±0.094**	0.010	**0.640±0.093**	0.075	**0.943±0.030**	0.005	**0.762±0.160**	0.086	**0.840±0.141**	0.047		

The average ± standard deviation for each metric is presented in terms of the type of data in the test dataset (WLI and NBI) and the combination of both (all), for each of the models. To have a reference point, the results obtained from physicians are shown too divided by specialists and residents. The table shows in which cases domain translation (DT) and unlabeled data (UD) were used during the training. The experiments to examine the impact of each of the branches (b_1 M, b_2, b_3) in the multi-input classifier were performed in a fully supervised (FS) way in order to analyze the effects only of the translations performed by the gan. The ablation results corresponding to branch b_1 is equivalent to the baseline (resnet-101) result since the inputs from csi-gan are not used. the cohen's kappa (CK) statistic is reported as an overall benchmark of the classifier.

TABLE VII
ABLATION RESULTS IN TERMS OF EACH OF THE CLASSES IN THE DATASET

name model	metric	HGC	p-val	LGC	p-val	NTL	p-val	NST	p-val
baseline (ResNet-101)	$Prec$	0.86±0.068	-	0.905±0.061	-	0.683±0.09	-	**0.941±0.036**	**-**
	Rec	**0.919±0.078**	-	0.849±0.130	-	0.869±0.084	-	0.865±0.081	-
	F-1	0.854±0.044	-	0.855±0.068	-	0.761±0.054	-	0.884±0.051	-
CSi-GAN-b_2	$Prec$	0.630±0.068	0.003	0.598±0.048	0.001	0.487±0.066	0.013	0.709±0.035	0.003
	Rec	0.669±0.064	0.005	0.708±0.099	0.151	0.300±0.082	0.003	0.770±0.059	0.254
	F-1	0.647±0.059	0.001	0.628±0.048	0.001	0.367±0.08	0.003	0.736±0.029	0.003
CSi-GAN-b_3	$Prec$	0.696±0.064	0.001	0.562±0.026	0.001	0.630±0.109	0.247	0.912±0.037	0.176
	Rec	0.649±0.060	0.002	0.660±0.050	0.032	0.560±0.060	0.002	0.865±0.013	0.731
	F-1	0.671±0.047	0.001	0.619±0.028	0.001	0.605±0.069	0.001	0.877±0.014	1.0
CSi-GAN	$Prec$	**0.919±0.029**	0.020	**0.943±0.036**	0.260	0.606±0.077	0.125	0.925±0.041	0.410
	Rec	0.885±0.031	0.319	0.868±0.070	0.230	0.880±0.081	0.972	0.824±0.062	0.723
	F-1	0.901±0.018	0.056	0.912±0.037	0.044	0.704±0.043	0.125	0.863±0.037	0.864
baseline semi-supervised	$Prec$	0.874±0.034	0.364	0.918±0.047	0.218	0.747±0.060	0.121	0.864±0.070	0.003
	Rec	**0.919±0.046**	0.953	0.840±0.042	0.791	0.840±0.078	0.233	0.865±0.030	0.360
	F-1	0.895±0.027	0.107	0.892±0.016	0.065	0.781±0.045	0.128	0.853±0.028	0.445
SeCSi-GAN	$Prec$	0.914±0.053	0.013	0.926±0.058	0.814	**0.778±0.060**	0.012	**0.941±0.075**	0.091
	Rec	**0.919±0.045**	0.877	**0.943±0.074**	0.013	**0.880±0.015**	0.072	**0.892±0.050**	0.009
	F-1	**0.914±0.040**	0.001	**0.922±0.044**	0.002	**0.800±0.109**	0.183	**0.895±0.040**	0.409

The average± is the standard deviation of each metric for each of the 4 classes. The experiments to examine the impact of each of the branches (b_1, b_2, b_3) in the multi-input classifier were performed in a fully supervised (FS) way in order to analyze the effects only of the translations performed by the gan. The ablation result corresponding to branch b_1 is equivalent to the baseline (resnet-101) result since the inputs from csi-gan are not used.

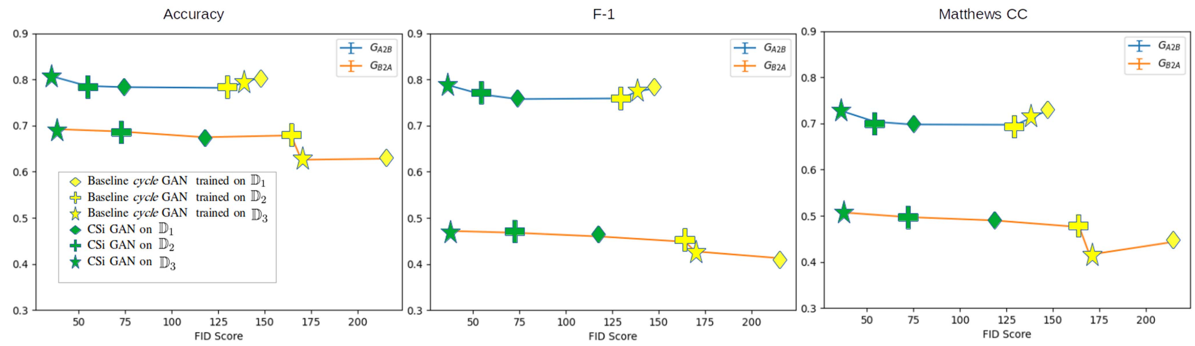


Fig. 7. Comparison of the different GAN models when used as backbone for training the multi-input classifier. The results are shown in terms of FID vs :ACC, F-1 score, and MCC.

performance. The metrics Acc, F-1 score, and MCC, obtained by training the multi-input classifier in a fully supervised using both *Cycle*GAN and CSi-GAN, are compared against the FID score for each of the translation networks. The results of this comparison are shown in Fig. 7. Even though it is easy to notice the gap in terms of the FID score between the generators from *Cycle*GAN and CSi-GAN, and the best classification metrics are obtained when using CSi-GAN with more data ($\mathbb{D}_3$), this improvement is minimum. Furthermore, *Cycle*GAN trained on $\mathbb{D}_2$ obtains similar metrics. The comparison against the classification metrics does not show a conclusive result and further research is needed to determine the correlations that could lead to best practices and parameter choices when training GAN models.

V. Conclusion

In this paper, we propose a novel semi-supervised learning GAN-based method to address the problem of endoscopic image classification in NBI and WLI imaging domains. The proposed method shows to be effective for a scenario where there is domain and class imbalance and in general, performs better than specialists and baseline methods. The use of this method leverages the use of unlabeled data in a domain different than the one where annotations exist, which is a very common case in biomedical data where annotated data is limited. This could ease the transition to clinical practice and its implementation for computer-aided BC diagnosis. The results obtained also show that the quality of the synthetic images generated with the proposed method is good enough to deceive clinical experts. Nevertheless, additional research needs to be carried out to find accurate metrics to assess the quality of generated images objectively and to determine to which point it might be related to the classification performances.

Future work includes further validation of multi-center data, as well as the acquisition of data from other imaging domains which could help to assess better the generalization of the method, and the development of lesion detection methods that could differentiate specific image regions that correspond to the lesion and non-lesion tissue. By making available our dataset we hope to encourage further research in the field that could motivate the clinical translation of endoscopic image classification.

Compliance with Ethical Standards

Ethical Approval: The proposed study is a retrospective study. No personal data was recorded. The collection of data was in accordance with the ethical standards of the Istituto Europeo di Oncologia and with the 1964 Helsinki declaration, revised in 2000. All the subjects involved in this research were informed and agreed to data treatment before the intervention.

Informed consent: Written informed consent was obtained from all patients included in the study.

References

[1] R. L. Siegel et al., “Cancer statistics,” 2021, *CA: A Cancer J. Clinicians*, vol. 71, no. 1, pp. 7–33, 2021.

[2] O. Sanli et al., “Bladder cancer,” *Nature Rev. Dis. Primers*, vol. 3, no. 1, pp. 1–19, 2017.

[3] K. C. DeGeorge et al., “Bladder cancer: Diagnosis and treatment,” *Amer. Fam. Physician*, vol. 96, pp. 507–514, 2017.

[4] R. Ball, “Pathology and genetics of tumours of the urinary system and male genital organs,” *Histopathology*, vol. 46, no. 5, pp. 586–586, 2005.

[5] M. C. Hall et al., “Guideline for the management of nonmuscle invasive bladder cancer (stages Ta, T1, and Tis): 2007 update,” *J. Urol.*, vol. 178, no. 6, pp. 2314–2330, 2007.

[6] H. W. Herr, “Narrow-band imaging evaluation of bladder tumors,” *Curr. Urol. Rep.*, vol. 15, no. 4, pp. 1–7, 2014.

[7] B. C. Jeong, “Recent technological advances in cystoscopy for the detection of bladder cancer,” in *Bladder Cancer*, J. H. Ku, Ed., Cambridge, MA, USA: Academic Press, 2018, ch. 10, pp. 135–144.

[8] Y. Y. Hui et al., “Wide-field imaging and flow cytometric analysis of cancer cells in blood by fluorescent nanodiamond labeling and time gating,” *Sci. Rep.*, vol. 4, no. 1, pp. 1–7, 2014.

[9] Z. Ye et al., “A comparison of NBI and WLI cystoscopy in detecting non-muscle-invasive bladder cancer: A prospective, randomized and multi-center study,” *Sci. Rep.*, vol. 5, no. 1, pp. 1–6, 2015.

[10] R. Chou et al., “Comparative effectiveness of fluorescent versus white light cystoscopy for initial diagnosis or surveillance of bladder cancer on clinical outcomes: Systematic review and meta-analysis,” *J. Urol.*, vol. 197, no. 3, pp. 548–558, 2017.

[11] R. J. Sylvester et al., “Predicting recurrence and progression in individual patients with stage ta T1 bladder cancer using EORTC risk tables: A combined analysis of 2596 patients from seven EORTC trials,” *Eur. Urol.*, vol. 49, no. 3, pp. 466–477, 2006.

[12] A. Nogueira-Rodríguez et al., “Deep neural networks approaches for detecting and classifying colorectal polyps,” *Neurocomputing*, vol. 423, pp. 721–734, 2021.

[13] K. Pogorelov et al., “Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection,” in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 164–169.

[14] P. Mesejo et al., “Computer-aided classification of gastrointestinal lesions in regular colonoscopy,” *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2051–2063, Sep. 2016.

[15] Y. Kominami et al., "Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy," *Gastrointestinal Endoscopy*, vol. 83, pp. 643–649, 2016.

[16] J. Xu et al., "Deep reconstruction-recoding network for unsupervised domain adaptation and multi-center generalization in colonoscopy polyp detection," *Comput. Methods Prog. Biomed.*, vol. 214, 2022, Art. no. 106576.

[17] E. Shkolyar et al., "Augmented bladder tumor detection using deep learning," *Eur. Urol.*, vol. 76, no. 6, pp. 714–718, 2019.

[18] R. Yang et al., "Automatic recognition of bladder tumours using deep learning technology and its clinical application," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 17, 2020, Art. no. e2194.

[19] A. Ikeda et al., "Support system of cystoscopic diagnosis for bladder cancer based on artificial intelligence," *J. Endourol.*, vol. 34, no. 3, pp. 352–358, 2020.

[20] N. Ali et al., "Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. e2194.

[21] M. Li et al., "Multi-domain few-shot image recognition with knowledge transfer," *Neurocomputing*, vol. 442, pp. 64–72, 2021.

[22] C. A. Lingley-Papadopoulos et al., "Computer recognition of cancer in the urinary bladder using optical coherence tomography and texture analysis," *Proc. SPIE*, vol. 13, 2008, Art. no. 024003.

[23] G. Csurka, "A comprehensive survey on domain adaptation for visual applications," *Domain Adapt. Comput. Vis. Appl.*, pp. 1–35, 2017.

[24] M. Misawa et al., "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database(with video)," *Gastrointestinal Endoscopy*, vol. 93, no. 4, pp. 960–967, 2021.

[25] K. Pogorelov et al., "Efficient disease detection in gastrointestinal videos–global features versus neural networks," *Multimedia Tools Appl.*, vol. 76, pp. 22493–22525, 2017.

[26] S. Nadeem et al., "Ensemble of texture and deep learning features for finding abnormalities in the gastro-intestinal tract," in *Proc. Int. Conf. Comput. Collective Intell.*, 2018, pp. 469–478.

[27] L. F. Sánchez-Peralta et al., "Piccolo white-light and narrow-band imaging colonoscopic dataset: A performance comparative of models and datasets," *Appl. Sci.*, vol. 10, no. 23, 2020, Art. no. 8501.

[28] J. Ahmad et al., "Endoscopic image classification and retrieval using clustered convolutional features," *J. Med. Syst.*, vol. 41, no. 12, pp. 1–12, 2017.

[29] S. Ali et al., "Additive angular margin for few shot learning to classify clinical endoscopy images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2020, pp. 494–503.

[30] M. R. Struyvenberg et al., "A computer-assisted algorithm for narrow-band imaging-based tissue characterization in Barrett's esophagus," *Gastrointestinal Endoscopy*, vol. 93, no. 1, pp. 89–98, 2021.

[31] S. Mohapatra et al., "Wavelet transform and deep convolutional neural network-based smart healthcare system for gastrointestinal disease detection," *Interdiscipl. Sci.: Comput. Life Sci.*, vol. 13, no. 2, pp. 212–228, 2021.

[32] S. Li et al., "Adaptive aggregation with self-attention network for gastrointestinal image classification," *IET Image Process.*, vol. 16, pp. 2384–2397, 2022.

[33] I. Lorencin et al., "On urinary bladder cancer diagnosis: Utilization of deep convolutional generative adversarial networks for data augmentation," *Biology*, vol. 10, no. 3, 2021, Art. no. 175.

[34] A. Rau et al., "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 7, pp. 1167–1176, 2019.

[35] R. J. Chenet al., "Slam endoscopy enhanced by adversarial depth prediction," in *Proc. KDD Workshop Appl. Data Sci. Healthcare*, 2019.

[36] S. Mathew et al., "Augmenting colonoscopy using extended and directional cyclegan for lossy image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4696–4705.

[37] S. Lin et al., "LC-GAN: Image-to-image translation based on generative adversarial network for endoscopic images," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 2914–2920.

[38] L. Sharan et al., "Mutually improved endoscopic image synthesis and landmark detection in unpaired image-to-image translation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 127–138, Jan. 2022.

[39] A. Marzullo et al., "Towards realistic laparoscopic image generation using image-domain translation," *Comput. Methods Prog. Biomed.*, vol. 200, 2021, Art. no. 105834.

[40] B. Yoo et al., "Joint learning of generative translator and classifier for visually similar classes," *IEEE Access*, vol. 8, pp. 219160–219173, 2020.

[41] Z. Zhang et al., "Joint optimization of CycleGAN and CNN classifier for detection and localization of retinal pathologies on color fundus photographs," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 115–126, Jan. 2022.

[42] S. Mabu et al., "Semi-supervised CycleGAN for domain transformation of chest CT images and its application to opacity classification of diffuse lung diseases," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 11, pp. 1925–1935, 2021.

[43] L. Cai et al., "A review of the application of deep learning in medical image classification and segmentation," *Ann. Transl. Med.*, vol. 8, no. 11, 2020, Art. no. 713.

[44] Q. Xie et al., "Self-training with noisy student improves imagenet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10687–10698.

[45] A. Odena, "Semi-supervised learning with generative adversarial networks," in *Proc. Workshop Data-Efficient Mach. Learn.*, 2016.

[46] W. Du et al., "Improving the classification performance of esophageal disease on small dataset by semi-supervised efficient contrastive learning," *J. Med. Syst.*, vol. 46, no. 1, pp. 1–13, 2022.

[47] M. Golhar et al., "Improving colonoscopy lesion classification using semi-supervised deep learning," *IEEE Access*, vol. 9, pp. 631–640, 2020.

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

[49] H. Shi et al., "Semi-supervised learning via improved teacher-student network for robust 3D reconstruction of stereo endoscopic image," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4661–4669.

[50] T. Salimans et al., "Improved techniques for training gans," *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[51] Z. Xue, "Semi-supervised convolutional generative adversarial network for hyperspectral image classification," *IET Image Process.*, vol. 14, no. 4, pp. 709–719, 2020.

[52] W. Li et al., "Semi-supervised learning using adversarial training with good and bad samples," *Mach. Vis. Appl.*, vol. 31, no. 6, pp. 1–11, 2020.

[53] L. Wang et al., "CCS-GAN: A semi-supervised generative adversarial network for image classification," *Vis. Comput.*, vol. 38, no. 6, pp. 2009–2021, 2022.

[54] S. Zhao et al., "Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions," in *Proc. AAAI Conf. Artif. Intell.*, pp. 2620–2627, vol. 33, no. 01, 2019.

[55] Y. Chen et al., "Cyclegan based data augmentation for melanoma images classification," in *Proc. 3rd Int. Conf. Artif. Intell. Pattern Recognit.*, 2020, pp. 115–119.

[56] M. Hammami et al., "Cycle GAN-based data augmentation for multi-organ detection in CT images via YOLO," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 390–393.

[57] C. Muramatsu et al., "Improving breast mass classification by shared data with domain transformation using a generative adversarial network," *Comput. Biol. Med.*, vol. 119, 2020, Art. no. 103698.

[58] Z. Xu et al., "Semi-supervised attention-guided cycleGAN for data augmentation on medical images," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 563–568.

[59] J.-Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[60] Z. Wang et al., "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[61] J. F. Lazo et al., "A transfer-learning approach for lesion detection in endoscopic images from the urinary tract," 2021, *arXiv:2104.03927*.

[62] A. Borji, "Pros and cons of GAN evaluation measures," *Comput. Vis. Image Understanding*, vol. 179, pp. 41–65, 2019.

[63] M. Heusel et al., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[64] D. Bashkirova et al., "Adversarial self-defense for cycle-consistent gans," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[65] A. Saif et al., "Abnormality detection in musculoskeletal radiographs using capsule network," *IEEE Access*, vol. 7, pp. 81494–81503, 2019.