Computer Methods and Programs in Biomedicine 234 (2023) 107511

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

The role of unpaired image-to-image translation for stain color normalization in colorectal cancer histology classification

Nicola Altini[a,1,*], Tommaso Maria Marvulli[c,1], Francesco Alfredo Zito[d], Mariapia Caputo[e], Stefania Tommasi[e], Amalia Azzariti[c], Antonio Brunetti[a,b], Berardino Prencipe[a], Eliseo Mattioli[d,1], Simona De Summa[e,1], Vitoantonio Bevilacqua[a,b,1]

[a] Department of Electrical and Information Engineering (DEI), Polytechnic University of Bari, Via Edoardo Orabona, 4, Bari 70126, Italy
[b] Apulian Bioengineering srl, Via delle Violette, 14, Modugno 70026, Italy
[c] Laboratory of Experimental Pharmacology, IRCCS Istituto Tumori "Giovanni Paolo II", Via O. Flacco, 65, Bari 70124, Italy
[d] Pathology Department, IRCCS Istituto Tumori "Giovanni Paolo II", Via O. Flacco, 65, Bari 70124, Italy
[e] Molecular Diagnostics and Pharmacogenetics Unit, IRCCS Istituto Tumori "Giovanni Paolo II", Via O. Flacco, 65, Bari 70124, Italy

ARTICLE INFO

Article history:
Received 30 October 2022
Revised 14 March 2023
Accepted 25 March 2023

Keywords:
Colorectal cancer
Generative adversarial network
Stain color normalization
Computer-aided diagnosis

ABSTRACT

Background: Histological assessment of colorectal cancer (CRC) tissue is a crucial and demanding task for pathologists. Unfortunately, manual annotation by trained specialists is a burdensome operation, which suffers from problems like intra- and inter-pathologist variability. Computational models are revolutionizing the Digital Pathology field, offering reliable and fast approaches for challenges like tissue segmentation and classification. With this respect, an important obstacle to overcome consists in stain color variations among different laboratories, which can decrease the performance of classifiers. In this work, we investigated the role of Unpaired Image-to-Image Translation (UI2IT) models for stain color normalization in CRC histology and compared to classical normalization techniques for Hematoxylin-Eosin (H&E) images.

Methods: Five Deep Learning normalization models based on Generative Adversarial Networks (GANs) belonging to the UI2IT paradigm have been thoroughly compared to realize a robust stain color normalization pipeline. To avoid the need for training a style transfer GAN between each pair of data domains, in this paper we introduce the concept of training by exploiting a meta-domain, which contains data coming from a wide variety of laboratories. The proposed framework enables a huge saving in terms of training time, by allowing to train a single image normalization model for a target laboratory. To prove the applicability of the proposed workflow in the clinical practice, we conceived a novel perceptive quality measure, which we defined as Pathologist Perceptive Quality (PPQ). The second stage involved the classification of tissue types in CRC histology, where deep features extracted from Convolutional Neural Networks have been exploited to realize a Computer-Aided Diagnosis system based on a Support Vector Machine (SVM). To prove the reliability of the system on new data, an external validation set composed of $N = 15{,}857$ tiles has been collected at IRCCS Istituto Tumori "Giovanni Paolo II".

Results: The exploitation of a meta-domain consented to train normalization models that allowed achieving better classification results than normalization models explicitly trained on the source domain. PPQ metric has been found correlated to quality of distributions (Fréchet Inception Distance — FID) and to similarity of the transformed image to the original one (Learned Perceptual Image Patch Similarity — LPIPS), thus showing that GAN quality measures introduced in natural image processing tasks can be linked to pathologist evaluation of H&E images. Furthermore, FID has been found correlated to accuracies of the downstream classifiers. The SVM trained on DenseNet201 features allowed to obtain the highest classification results in all configurations. The normalization method based on the fast variant of CUT (Contrastive Unpaired Translation), FastCUT, trained with the meta-domain paradigm, allowed to achieve the best classification result for the downstream task and, correspondingly, showed the highest FID on the classification dataset.

* Corresponding author.
E-mail address: nicola.altini@poliba.it (N. Altini).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2023.107511

Conclusions: Stain color normalization is a difficult but fundamental problem in the histopathological setting. Several measures should be considered for properly assessing normalization methods, so that they can be introduced in the clinical practice. UI2IT frameworks offer a powerful and effective way to perform the normalization process, providing realistic images with proper colorization, unlike traditional normalization methods that introduce color artifacts. By adopting the proposed meta-domain framework, the training time can be reduced, and the accuracy of downstream classifiers can be increased.

1. Introduction

Colorectal cancer (CRC) is the second cause of death for cancer with mortality reaching almost 35% [1]. In the last few years, new targeted therapies have been developed gaining significant improvement in clinical outcomes for several malignancies [62]. To date, a shift from tumor cells to the tumor microenvironment (e.g., for immunotherapic treatments) highlighted the importance to know cell-cell interaction in the context of tissue morphology. As an example, there is a growing interest in the knowledge of the spatial location of transcriptomic data. Thus, the segmentation of tissue types is required to better perform spatial analyses, e.g., for the selection of relevant regions of interest. Moreover, it is well-known that stroma-rich CRC has a poor prognosis [2], and a tissue segmentation pipeline could be helpful in prognosis prediction.

Image processing and Deep Learning (DL) techniques can be exploited for the automatic analysis of histological images, e.g., tissue type classification. The traditional workflow for realizing an image classifier is composed of several stages, i.e., preprocessing, feature extraction, dimensionality reduction, and classification, that can be obtained with models like Support Vector Machines (SVMs), Decision Trees (DTs), and Artificial Neural Networks (ANNs). DL-based workflows, instead, can enable end-to-end training of the models, easing complex steps, such as handcrafted feature extraction, and leading to performance improvement. These techniques can be exploited for developing Computer-Aided Diagnosis (CAD) systems which can enhance pathologists' workflows, reducing issues concerning intra- and inter-pathologist variability [3].

Nevertheless, a fundamental problem regarding the histopathological classification of images arises from the differences in colors between tissue samples from different institutions. Indeed, a complex protocol composed of several steps, namely: (i) collection and fixation, (ii) dehydration and clearing, (iii) paraffin embedding, (iv) microtomy, (v) staining, (vi) mounting, and (vii) digitalization, is required for Digital Pathology workflows [3]. Artifacts and differences among laboratories can be introduced at any of these stages [4].

Stain color normalization is therefore a pivotal pre-processing step for successfully deploying Deep Learning CAD frameworks in Digital Pathology setups [5,6]. Currently, the taxonomy of stain color normalization methods comprises: (a) global color normalization; (b) color normalization after stain separation; (c) color transfer with deep neural networks [4]. The first two approaches involve traditional image processing techniques, whereas the third harnesses the power of DL. Indeed, in recent years, the possibilities offered by Generative Adversarial Networks (GANs) [40], to effectively implement color transfer between histopathological images, are improving the performances of classification systems that can rely upon higher quality normalized images [7]. Other authors considered the problem of stain color normalization in nuclei segmentation pipelines [8], or for normalization of tissue of breast and prostate cancers [9,10], but most of these works limit their analysis to Conditional Generative Adversarial Networks (cGAN) [11], CycleGAN [12] and its variants, such as Residual Cycle-GAN [13].

The Image-to-Image Translation paradigm [11] can be thought as a general framework to tackle a variety of image analysis problems, such as segmentation, color normalization, reconstruction of original images from labels, and conversion from one source modality to another, among the others. This approach consists in training a conditional GAN [42] for translating images from a source domain to a target domain; however, such architectures need a dataset with paired images for setup and training. On the other hand, CycleGAN [12] and several subsequent works [43–49] focused on the idea to realize the image domain translation having only the domain as the label. This setting does not require paired data and has led to the concept of Unpaired Image-to-Image Translation (UI2IT). Such paradigm allows to construct datasets for normalization in a manner that is affordable for Digital Pathology laboratories, as paired image data is normally not available, especially when data come from two distinct institutions.

In this work, we aimed to realize a reliable pipeline for stain color normalization and tissue classification in H&E samples of patients with CRC. For the stain color normalization stage, five GANs [12,43,48,49] based on the UI2IT framework have been thoroughly compared. Furthermore, four traditional image processing normalization techniques [14–17] have been considered as baselines. In order to assess the feasibility of the proposed normalization methodology, an evaluation of the generated tiles has been realized by an expert pathologist, introducing a metric that we defined as Pathologist Perceptive Quality (PPQ). Afterwards, to realize the tissue classifier, three CNNs have been considered as feature extractor from tile normalized with the previously mentioned techniques. An SVM has been trained on top of deep features, in order to assess the classification accuracy of the developed CAD system.

Contrarily to what is usually done in stain color transfer, where a generative model is trained between each pair of domains, a meta-domain — The Cancer Genome Atlas (TCGA) — composed of the union of data coming from several laboratories, has been considered in place of the source domain in the training phase of the stain color normalization module. With the only need to eventually perform a double normalization at inference time, on both source and target classification domains, the proposed methodology has the advantage of avoiding the expensive process of training multiple GANs. The proposed meta-domain methodology has been compared to the standard approach for GAN-based stain color transfer, i.e., learning the stain transfer mapping directly from the source domain to the target domain.

Summarizing, in this work, we added the following innovative contributions: (i) an extensive comparison of normalization techniques in order to assess the most reliable ones for validating tissue classifiers on data coming from different laboratories; (ii) an investigation of features extracted from deep CNN architectures for CRC tissue classification; (iii) an evaluation method for assessing the quality of generated tiles from expert pathologists, resulting in the conceptualization of a novel perceptive metric, PPQ; (iv) a setup for UI2IT stain color normalization which does not require the need for training style transfer GANs between every pair of data domains, via the exploitation of a meta-domain; (v) a collec-

tion of a validation cohort of samples enrolled at the IRCCS Istituto Tumori Bari "Giovanni Paolo II", resulting in a publicly available dataset of 10 WSIs and $N = 15{,}857$ annotated tiles [19].

The remainder of the paper is structured as follows. Section 2 details related works in both colorectal cancer tissue classification and stain color normalization, encompassing techniques which are traditionally adopted for H&E images and those belonging to the UI2IT framework. Section 3 presents materials and methods adopted for this research. Firstly, the employed and collected datasets are described. The experimental UI2IT setting, which features a meta-domain source dataset, is introduced. Evaluation metrics for assessing the quality of generated normalized images are described, with specific considerations for histopathological scenarios. Section 4 presents the experimental results, both for the quality of normalized images and for the classification task. Results are then discussed in Section 5, where also limitations and directions for future works are presented. Finally, conclusions are portrayed in Section 6.

2. Related works

2.1. Colorectal cancer tissue classification

The issue of classifying epithelium and stroma from digitized tumor tissue microarrays (TMAs) has been considered by Linder et al. in 2012 [20]. In the feature extraction phase, the authors took advantage of LBP (Local Binary Patterns) and LBP/C, where C is a contrast measure. Other features considered were Gabor filtered images and Haralick texture features [21,22]. Employing an SVM classifier, the authors stated that the LBP/C-based is the best one, with an AUC (Area Under the Curve) ROC (Receiver Operating Characteristic) of 0.995.

Kather et al. considered a multi-class tissue classification in the domain of colorectal cancer histopathology [23]. During the feature extraction stage, they considered several categories of features, after having transformed the original color images into gray-scale ones: histogram features, of both lower-order and higher-order; LBP; Gabor filters; gray-level co-occurrence matrix (GLCM); perception-like features. For the classification step, the authors investigated four classifiers: decision trees, linear SVM, radial-basis function SVM, and 1-nearest neighbor. The same feature set can obtain higher results by exploiting red channel versions of images instead of gray-scale ones [24], even though this observation holds mainly for unnormalized images. Kather et al. also exploited the capabilities of CNNs for the sake of classifying CRC Hematoxylin-Eosin (H&E) histopathology images of TCGA composed of 862 whole slide images (WSIs) [25].

Even though Ciompi et al. [6] claimed the importance of stain color normalization for CRC tissue classification, posing the focus on classical techniques [15,16,18], from the works analyzed in this section, it emerges that no systematic investigation of recent methods based on UI2IT has been carried out in this context.

Indeed, most of these studies are tailored to discover the most efficient features or classification architecture for the task in hand, without a proper consideration of the pre-processing steps such as stain color normalization.

2.2. H&E normalization

Histopathology involves a manual staining procedure for preparing tissues prior to microscopic imaging for diagnosis. This is a non-standardized procedure which may cause considerable variability in the color characteristics of tissue samples from different laboratories; this can occur due to inconsistent tissue staining, different color responses to distinct scanners, or differences in raw materials and stain manufacturing techniques. Stain color variation degrades the performance of CAD systems. In the presence of severe color variation in histopathological images, stain color normalization, which is achieved by removing the stains for visual enhancement, is a common practice.

Among the most popular image processing methods, it is worth mentioning the works of Reinhard et al. [14], Macenko et al. [15], Khan et al. [16], and Vahadane et al. [17]. These works are usually considered as reference methods when authors propose novel methods for stain color normalization [7,11,27,31,32]. Nevertheless, important limitations of these methods include the fact that they require H&E staining, the need of exploiting a template patch for fitting the stain distributions [13], and the introduction of color artifacts. On the other hand, the proposed stain normalization methodology, based on UI2IT, can be applied to every type of staining, and are not restricted to H&E images only.

For more details about normalization techniques and preprocessing procedures, the interested reader is referred to dedicated surveys [4,33] or to the original papers mentioned before.

2.3. Stain color normalization for colon histological tissue

Several DL-based methods have been proposed to tackle the stain color normalization problem for histopathological tissues including the colon [8,13,26–28].

Bentaieb et al. proposed a stain transfer-based approach for stain color normalization [26]. The authors designed a discriminative image analysis network that has the capability to relocate stains between different datasets. Their architecture is composed of a generative network devoted to learning both dataset-specific staining properties and image-specific color transformation, and a task-specific network which is exploited for the downstream task (as segmentation or classification). Their model can be trained end-to-end by exploiting a multi-objective loss. The authors' conclusion states that their model is capable of improving the results, both for what concerns the quality of normalized images, and the accuracy of the networks for the downstream tasks, over various baselines.

Pontalba et al. evaluated the impact of several existing methodologies for stain color normalization in a setup for nuclei segmentation [8]. The considered normalization techniques comprise histogram specification, color transfer, stain specific color transfer, spectral matching, and CycleGAN. The authors also considered several measures for assessing the quality of normalized images, besides evaluating the results of the downstream segmentation task. To enhance the CycleGAN capabilities, de Bel et al. proposed an improvement over the base model, with the embodiment of residual learning, devising an architecture that they defined as Residual CycleGAN [13]. The authors compare the performance of their stain normalization approach, also with respect to data augmentation, to prove the robustness of the downstream segmentation networks. The considered downstream applications include segmentation from colon and kidney tissue samples.

Shen et al. noted that the transformation induced by GANs can cause information loss, or suffer from problems such as mode collapse, damaging results for the subsequent diagnostic task [27]. To solve this problem, they devised a contrastive learning method with a color-variation constraint, to retain the recognizable phenotypic features when using a GAN for stain color normalization. Self-supervised learning allows to cluster discriminative tissue patches among several types of tumors.

Kausar et al. introduced a deep model, which they defined as Stain Acclimation Generative Adversarial Network (SA-GAN), which has an architecture that comprises one generator and two discriminators [28]. As usual, the purpose of the generator is to transform images from the source domain to the target domain. The two dis-

criminators, instead, have different roles: the first one enforces the generated images to retain the color patterns of the target domain, whereas the second one enforces the generated images to maintain the structure contents of the source domain.

With respect to the work summarized for the stain color normalization, we observe that: (i) other authors have not applied these techniques for our application, that is multiclass CRC tissue classification, as presented in Section 3; (ii) other authors did not always found a suitable way to include pathologists evaluations or to correlate them with existing metrics for assessing GAN-generated image quality or accuracy of the downstream task; (iii) other authors have not explored the whole UI2IT framework, but rather focused on single models, especially those similar to CycleGAN [8,13] or contrastive learning [27].

2.4. Unpaired image-to-image translation

Stain normalization of histopathological images performed with GANs is gaining much attention recently [7], with a particular focus on CycleGAN [12] and its variants [13,34]. Nonetheless, many other algorithms belonging to the UI2IT framework have not been explored for the stain color transfer task.

In this subsection, we summarized and categorized relevant works based on GANs for UI2IT, to explain the role of UI2IT in the normalization scenario for histopathological images. In the following, S and T denote the source and target domain, whereas s and t are instances of the two domains.

The interested readers may find useful information also in surveys concerning about adversarial-learning-based I2IT [35–37] or GAN applications, techniques for training, and architectures [38–41].

2.4.1. Cycle consistency-based

In UI2IT setups, *cycle-consistency* is the most widely adopted method for imposing association [43]. This paradigm is grounded on the concept of retrieving also the reverse mapping from the target domain back to the source one. Furthermore, it enforces a check that a sample input image can be reconstructed. Among the most well-known architectures which fall into this category, it is worth mentioning CycleGAN [12], DualGAN [44], and DiscoGAN [45].

CycleGAN overcomes the limitations of the paired I2IT framework by learning a mapping G_{ST}: $S \rightarrow T$ in a way that $G_{ST}(S)$ is indistinguishable from T exploiting an adversarial loss. To avoid the issues coming from the under-constrained mapping, at the same time, an inverse mapping G_{TS}: $T \rightarrow S$ is also learned. In this way, the cycle consistency loss can enforce $G_{TS}(G_{ST}(S)) \approx S$ and $G_{ST}(G_{TS}(T)) \approx T$.

2.4.2. One-sided translation

Instead of enforcing cycle-consistency, it is possible to promote relationships belonging in the input to be similarly reflected in the output. For instance, patches which are perceptually similar inside an input image should retain their proximity in the output. TraVeLGAN [46], DistanceGAN [47], and GcGAN [48] allow one-way translation, so avoiding the need for a cycle-consistency. The problem is that they require relationships between full images, or with predefined distance functions.

While the cycle consistency framework needs to train two generators simultaneously, G_{ST} and G_{TS}, one-directed domain translation can be successfully achieved also by only preserving pairwise images' distances. An important limitation of both cycle consistency and distance preservation, is that they do not properly consider simple geometric transformations.

The idea of enforcing a geometry-consistency constraint in a UI2IT GAN framework comes from the work of Fu et al. [48]. According to the authors, considering a geometric transformation $f(\cdot)$, the images from the source domain should not be altered by the corresponding generators G_{ST} and $G_{S'T'}$, where S' and T' are the domains obtained by applying $f(\cdot)$ to S and T, respectively.

From a mathematical perspective, considering a random sample s from original domain S, a proper geometric transformation $f(\cdot)$, and its inverse $f^{-1}(\cdot)$, the geometry consistency constraint can be formulated as $f(G_{ST}(s)) \approx G_{S'T'}(f(s))$ and $f^{-1}(G_{S'T'}(f(s))) \approx G_{ST}(s)$. Since it is improbable that G_{ST} and $G_{S'T'}$ make errors in the same region, the generator models act as co-regulators for each other, thanks to the geometry consistency constraint, thus improving over mistakes in local zones of their relative translations.

It is worth noting that the one-sided translation is particularly interesting for stain color normalization, since one usually wants to translate tiles from the original domain to the target domain, and therefore there is no need to exploit both generators.

2.4.3. Patchwise contrastive learning approaches

CUT and FastCUT [43] have been proposed to solve the limitations encountered in relationship preservation-based architectures, by replacing cycle-consistency with the possibility of learning a cross-domain similarity function by maximizing mutual information between corresponding patches from images belonging to the source and target domains, without the need to depend on some predefined distance.

The architecture proposed by Park et al. [43] enforces positive (related) patches to map to nearby points in the learned feature space, if compared to negative (unrelated) patches coming from the dataset. This framework allows one-sided translation in UI2IT setups, resulting in a greater quality and less training time compared to cycle-consistency-like approaches.

Mathematically, we can define: the query $v \in \mathbb{R}^K$, the positive $v^+ \in \mathbb{R}^K$, and N negative samples $v^- \in \mathbb{R}^{N \times K}$. In these definitions, v, v^+, v_i^- are K-dimensional vectors, and v_i^- is the i-th vector from the matrix of negatives v^-. The cross-entropy loss can then be calculated, defining the probability of a positive example to be chosen over negatives, as reported in Eq. (1).

$$l\left(v, v^+, v^-\right) = -log\left[\frac{exp\left(v \cdot \frac{v^+}{\tau}\right)}{exp\left(v \cdot \frac{v^+}{\tau}\right) + \sum_{i=1}^{N} exp\left(v \cdot \frac{v_i^-}{\tau}\right)}\right] \tag{1}$$

τ is a temperature defined as equal to 0.07 by the original authors. The goal of the CUT framework is to relate source and target image patches. In the considered context, the query concerns the target, whereas the positive and negative samples concern corresponding and noncorresponding source patches, respectively.

In order to enforce this relationship, the authors selected J layers with whom they encoded the input images and passed them to a two-layers MLP H_j, creating a stack of features $\{z_j\}_J = \{H_j(G_{enc}^j(s))\}_J$, where G_{enc}^j is the output of the j-th layer. In a similar way, they also encoded the output image in $\{\hat{z}_j\}_J = \{H_j(G_{enc}^j(G(s)))\}_J$. The objective is to match the corresponding source-target patches for each layer and use the other ones from the source image as the negatives.

Then, they defined, for layer j, the features of the positive patch as $z_j^q \in \mathbb{R}^{C_j}$ and the features of the negative patches as $z_j^{Q\setminus q} \in \mathbb{R}^{(Q_j-1)\times C_j}$, where $j \in \{1, 2, ..., J\}$ indexes the selected layer and $q \in \{1, 2, ..., Q_j\}$ indexes the spatial location. The number of spatial locations and the number of channels for layer j are referred to as Q_j and C_j, respectively.

The loss is calculated as reported in Eq. (2).

$$L_{PatchNCE}(G, H, S) = E_{s\sim S}\left[\sum_{j=1}^{J}\sum_{q=1}^{Q_j} l\left(\hat{z}_j^q, z_j^q, z_j^{Q\setminus q}\right)\right] \tag{2}$$

The final objective function can be defined as portrayed in Eq. (3):

$$L_{GAN}(G, D, S, T) + \lambda_S L_{PatchNCE}(G, H, S) + \lambda_T L_{PatchNCE}(G, H, T) \quad (3)$$

where $L_{PatchNCE}(G,H, T)$ is the identity loss that enforces the generator to avoid unnecessary changes. The configuration with $\lambda_S = 1$ and $\lambda_T = 1$ corresponds to CUT, whereas the configuration with $\lambda_S = 10$ and $\lambda_T = 0$ is referred to as FastCUT.

AI-FFPE [49] is a modification of CUT that it adds some unique characteristics, like the presence of a Spatial Attention Block (SAB) in the architecture of the generator and a L_1 regularization factor in the objective function, that is calculated as shown in Eq. (4):

$$L_{GAN}(G, D, S, T) + \lambda_{reg} L_{reg}(G, S) + \lambda_S L_{PatchNCE}(G, H, S) + \lambda_T L_{PatchNCE}(G, H, T) \quad (4)$$

where $L_{reg}(G,S) = ||S - G(S)||_1$.

The authors of AI-FFPE claim that in the frozen section to FFPE (formalin-fixation and paraffin-embedding) translation task, their model outperforms generic image-translation networks. Particularly, they state that their modifications to the CUT architecture contribute to the artifact-correcting performance of their model.

3. Materials and methods

3.1. Datasets

Two kinds of datasets have been collected for this research: datasets for stain color normalization and for multi-class classification. All datasets contain histopathological image tiles belonging to WSIs of patients diagnosed with CRC.

For what concerns stain color normalization, the considered datasets are: a dataset introduced by Kather et al. [29] (SD1) and a local dataset collected at IRCCS Istituto Tumori "Giovanni Paolo II" (SD2). The SD1 dataset is composed of image tiles coming from 604 CRC WSIs in the TCGA database, while the SD2 dataset is composed of image tiles coming from 58 WSIs. All tiles have dimensions of 512 × 512 px at 0.5 µm/px. SD1 and SD2 have been used to construct two tile-level datasets, each with a training set of $N = 100{,}000$ tiles and a test set of $N = 50{,}000$ tiles.

With respect to the multi-class classification, the considered datasets are: a dataset introduced by Kather et al. [25,30] (CDT) and a local dataset collected at IRCCS Istituto Tumori "Giovanni Paolo II" (CDV).

The CDT dataset [25,30] is composed of $N = 100{,}000$ image tiles from H&E histological tissue of humans with CRC, subdivided into nine tissue classes. The size of the images is 224 × 224 pixels, which correspond to 112 × 112 µm^2.

The CDV dataset is composed of $N = 15{,}857$ tiles coming from 10 WSIs. The tiles have dimension 224 × 224 pixels which correspond to 116 × 116 μm^2. The dataset has been annotated by an expert pathologist. We made our dataset publicly available [19], to ease the development and comparison of computational techniques for CRC histological image analysis.

Both the datasets have been classified into the following seven classes, as done in our precedent work [24]: *TUM* – tumor epithelium; *MUSC_STROMA* – the union of SIMPLE_STROMA, encompassing smooth muscle, tumor stroma and extra-tumor stroma, and COMPLEX_STROMA, consisting of stroma or smooth muscle containing single tumor cells and/or few, non-aggregated immune cells; *LYM* – lymphoid follicles and other immune-cell conglomerates; *DEBRIS_MUCUS* – hemorrhage, mucus and necrosis; *NORM* – normal mucosa; *ADI* – adipose tissue; *BACK* – background.

To accomplish this categorization for the CDT dataset, the *DEBRIS_MUCUS* class has been built by combining the DEB and MUC classes; instead, the *MUSC_STROMA* class has been constructed by fusing the MUS and STR classes. The CDT dataset obtained with this procedure is composed of $N = 77{,}805$ tiles, considering that only half of the images of the merged classes have been retained, for maintaining class balancing during training.

For both the locally collected datasets, namely, *SD2* and the *CDV*, the Institutional Ethics Committee of the IRCCS Istituto Tumori "Giovanni Paolo II" approved the study (Prot. n. 780/CE).

Fig. 1 portrays frequency distribution and example tiles from normalization and classification datasets. Instead, color artifacts introduced by the classical normalization methods presented in Section 2.2 are shown in Fig. 2.

3.2. Stain normalization with a meta-domain

In the standard pipelines developed for stain color normalization exploiting UI2IT frameworks, two domains are usually considered: a source domain S and a target domain T. The objective is to transform images of the source domain to the distribution of the target domain. Therefore, a generator, to learn a map from S to T, namely $G_{ST}(S) \approx T$, is the outcome of the training stage for stain color transfer. The inverse mapping is referred to as G_{TS}, and is required only for the cycle-consistency-based methods. On the other hand, traditional normalization methods exploit a reference tile, R.

Instead, in the framework proposed for the CAD for CRC classification, we considered three domains: the meta-domain M, the source domain S, and the target domain T.

In our application, M is a composition of multiple subdomains S_1,..., S_n, covering a wide variety of stain color conditions, so that by learning $G_{MT}(M) \approx T$, we are capable to approximately map $G_{MT}(S) \approx T$. In the proposed configuration, the results can be further improved from performing a double normalization, i.e., $G_{MT}(S)$ and $G_{MT}(T)$. Indeed, we note that $G_{MT}(S)$ is more similar to $G_{MT}(T)$ than to T. This is in contrast to the usual way to perform stain transfer with GANs, where the baseline UI2IT consists in comparing distribution of images of $G_{ST}(S)$ with T.

When referring to our application, M is the domain of images coming from the TCGA (meta-domain), S is the domain of images coming from the training set for the classification (source domain), and T is the domain of images coming from our local cohort (target domain). The proposed framework may lead to improvements in generalization with respect to performing the traditional $G_{ST}(S)$.

3.3. Generated image tiles quality assessment

In this subsection, relevant metrics which can be used for assessing the quality of generated images are presented, so that both mathematical and perceptual evaluations can be done.

Quantitative evaluation of the quality of images generated by GANs is not an easy task, but different approaches have been proposed in literature. Common quality objective measures for image similarity include Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [50]. As noted by Zhang et al. [51], these metrics are shallow, simple functions, which cannot measure in a proper way the human perception. They proposed to realize *a perceptual distance* with deep features exploiting the VGG network, resulting in a Learned Perceptual Image Patch Similarity (LPIPS) metric. Unluckily, these measures can be exploited only if ground-truth images are available [35], so they are excellent candidates to assess results obtained with conditional GAN for paired I2IT tasks, but are not directly applicable in UI2IT settings, at least to assess adherence of the mapped image to the target domain. On the other hand, these measures can be considered to evaluate the introduction of artifacts between original images and their translated version.

Scores obtained exploiting Inception-v3 network pretrained on ImageNet, such as Inception Score (IS) [52], and Fréchet Inception

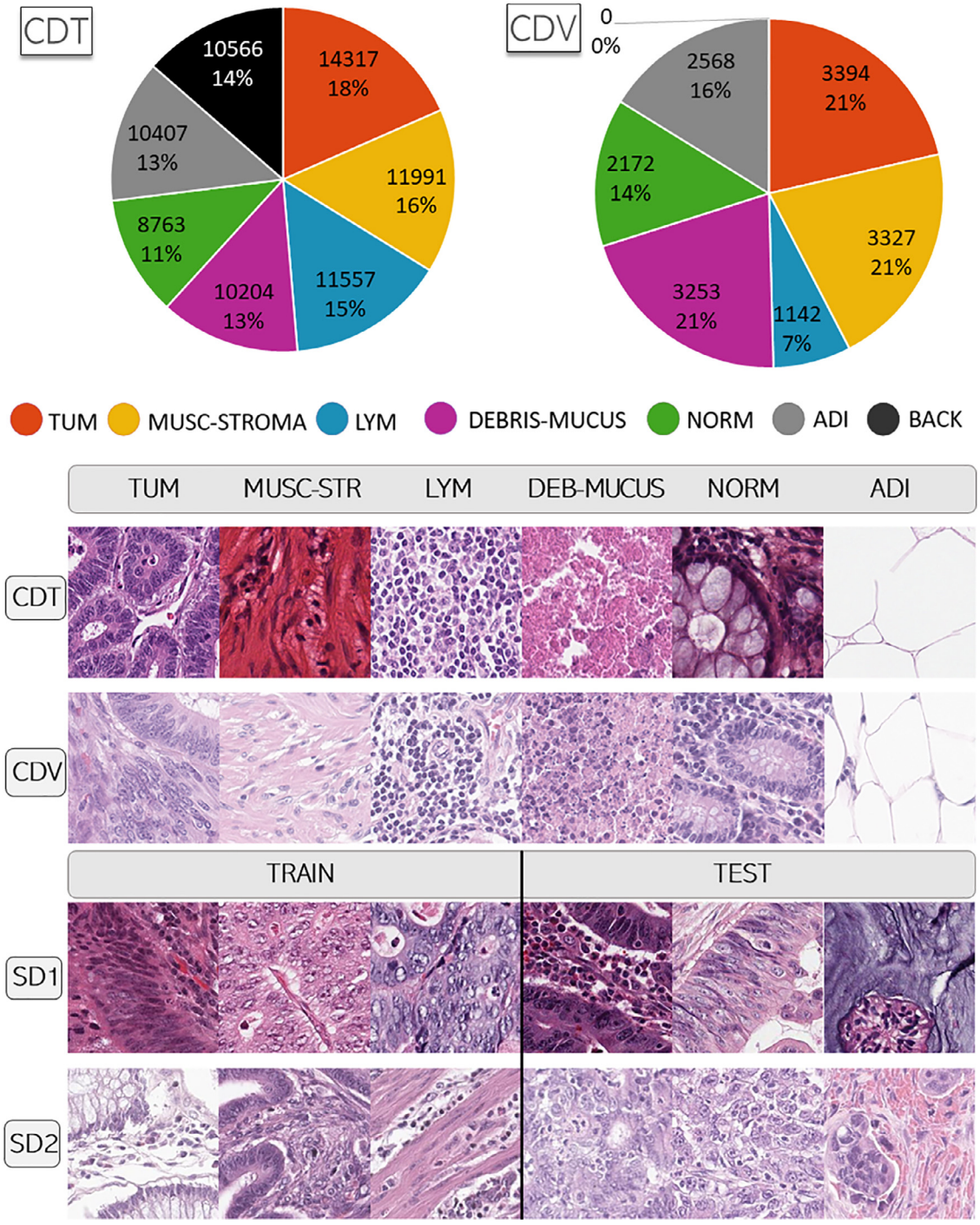

Fig. 1. Datasets exploited for the stain normalization and the downstream classification task. SD1 and SD2 are the datasets for stain normalization. CDT and CDV are the datasets for the classification task.

Distance (FID) [53] were introduced to overcome this issue and allow an assessment of realness and heterogeneity of generated images, from the point of view of feature distribution.

To determine the FID, two multivariate Gaussians are fitted on feature vectors obtained by embedding samples from the Inception network. Then, the Fréchet Distance [54] or the Wasserstein-2 distance [55] is calculated among these two gaussian distributions, as in Eq. (5):

$$FID(r,f) = \| \mu_r - \mu_f \|_2^2 + Tr\left(\Sigma_r + \Sigma_f - 2\left(\Sigma_r\Sigma_f\right)^{\frac{1}{2}}\right) \quad (5)$$

where μ_r, μ_f represent the mean of the real and fake generated sample feature vectors, respectively, and Σ_r, Σ_f represent the covariance matrix of the real and fake generated sample feature vectors, respectively.

In the research community, there is not a wide agreement on how to evaluate unpaired image-to-image translation frameworks [56]. Therefore, the considered experimental design involves the adoption of several metrics among the considered datasets.

To assess the reliability of the fake images generated via GANs, a novel perceptive quality measure has been introduced, which we have defined as PPQ. In detail, an image tile and the related

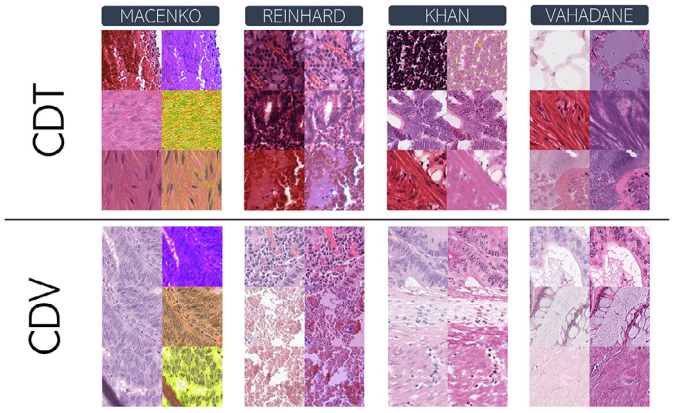

Fig. 2. Example of artifacts introduced by classical stain color normalization methods. (Top Row) Samples belonging to the training set for classification (CDT). (Bottom Row) Samples belonging to the test set for classification (CDV).

GAN-normalized versions of the same tile were shown to an expert pathologist, and for each normalization method the pathologist had to decide a discrete score between 1 and 4, where 1 represents an excellent quality image and 4 a bad one. In our experimental design, 200 image tiles were shown to the pathologist, for both the SD1 and SD2 datasets.

3.4. Experimental design

Given their importance as a benchmark, we decided to include the classical normalization methods [14–17] as baseline methods.

Among Cycle Consistency-based networks, due to its wide adoption in the stain color normalization domain [8–10,31], only CycleGAN has been included in the experiments. Amid the One-sided Translation networks, GcGAN has been included in our experiments, with the 90° clockwise rotation as geometric transformation, since it has been proven useful by the authors of the original paper. Among Patchwise Contrastive Learning approaches, the CUT, FastCUT and AI-FFPE models have been included in our analysis. All GAN models have been trained from scratch with a learning rate of 0.0002, for 5 epochs, with the Adam solver [57] and a batch size of 1. Input images were of size 256 × 256. No other preprocessing was applied to the images.

In order to perform a comprehensive comparison and analysis of normalization methods and downstream classifiers, we designed the experiments as explained in this section.

CDT and *CDV* refer to the training and test set for classification, respectively. $N(CDT)$ and $N(CDV)$ refer to their normalized version, that is $G_{MT}(CDT) = CDT_{MT}$ and $G_{MT}(CDV) = CDV_{MT}$ for the GAN normalization with the meta-normalization paradigm, $G_{ST}(CDT) = CDT_{ST}$ and $G_{ST}(CDV) = CDV_{ST}$ for the standard GAN normalization, and CDT_{RT} and CDV_{RT} for classical normalization, respectively.

*SD*1 and *SD*2 refer to the stain normalization datasets. $N(SD1)$ and $N(SD2)$ refer to their normalized version, that is $G_{MT}(SD1) = SD1_{MT}$ and $G_{MT}(SD2) = \mathrm{SD2_{MT}}$ for the GAN normalization, and $SD1_{RT}$ and $SD2_{RT}$ for the classical normalization, respectively. Images from *CDV* and *SD*2 belong to the *T* domain, images from *SD*1 belong to the *M* domain, and images from *CDT* belong to the *S* domain.

For what concerns normalization procedure:

(1) A reference tile has been used for the traditional normalization methods (Reinhard, Macenko, Khan, Vahadane). Both the *CDT* and *CDV* datasets image tiles have been normalized when used for classification. In order to assess normalized image distributions for these methods, FID has been calculated between:
   (a) $SD1_{RT}$ and $SD2_{RT}$,
   (b) CDT_{RT} and CDV_{RT}.
(2) In order to obtain a model capable of mapping *CDT* and *CDV* image tiles to the same target domain, all the GANs have been trained by exploiting *SD*1 and *SD*2. The following tests have been made in order to evaluate if a double normalization (that is, normalizing images by performing $G_{MT}(S)$ and $G_{MT}(T)$) would work better than a single normalization (i.e., normalizing images by performing $G_{MT}(S)$ only, while keeping *T* images unaltered), in our setting. In detail, to assess the quality of the generated image distributions, FID has been calculated between:
   (a) $SD1_{MT}$ and *SD*2,
   (b) $SD1_{MT}$ and $SD2_{MT}$,
   (c) CDT_{MT} and *CDV*,
   (d) CDT_{MT} and CDV_{MT}.
(3) In order to understand the quality of the normalized images with respect to the original ones, PSNR, SSIM, and LPIPS have been determined between:
   (a) $SD1_{RT}$ and *SD*1,
   (b) $SD2_{RT}$ and *SD*2,
   (c) $SD1_{MT}$ and *SD*1,
   (d) $SD2_{MT}$ and *SD*2.
(4) In order to understand the quality of the GAN-generated images, from the pathologist perspective, PPQ has been calculated considering:
   (a) $SD1_{MT}$ and *SD*1,
   (b) $SD2_{MT}$ and *SD*2.
(5) In order to understand how UI2IT models affect saturation of images, SSIM between saturation channels (in HSV color space) has been calculated between:
   (a) $SD1_{MT}$ and *SD*1,
   (b) $SD2_{MT}$ and *SD*2.

For what concerns the downstream classification, three CNNs have been exploited as feature extractor before training an SVM classifier. In detail, the employed SVM was a multi-class classification error-correcting output code (ECOC) model with one-vs-one coding design and Radial Basis Function (RBF) kernel. The three CNNs considered are DenseNet201 [58], InceptionV3 [59], and VGG16 [60]. The SVM model has been trained on features extracted from $N(CDT)$ and validated on features belonging to $N(CDV)$.

In particular, the following configurations have been compared for the classification:

(1) SVM trained on *CDT* and validated on *CDV*, with no normalization.
(2) SVM trained on CDT_{RT} and validated on CDV_{RT}, for the considered classical normalization techniques.
(3) SVM trained on CDT_{ST} and validated on CDV_{ST}, for the considered UI2IT GAN-based approaches, trained from source to target domain (baseline UI2IT).
(4) SVM trained on CDT_{MT} and validated on CDV_{MT}, for the considered UI2IT GAN-based approaches, trained from meta-domain to target domain (meta-domain UI2IT).

A detail of the experimental design is pictorially represented in Fig. 3.

4. Experimental results

The results obtained from the experiments described in Section 3.4 are presented in this section. In order to better characterize the two components of the developed pipeline, three subsections are delineated. The first one deals with the results of the stain color normalization, whereas the second one presents the results of the multi-class classification. Lastly, the third section de-

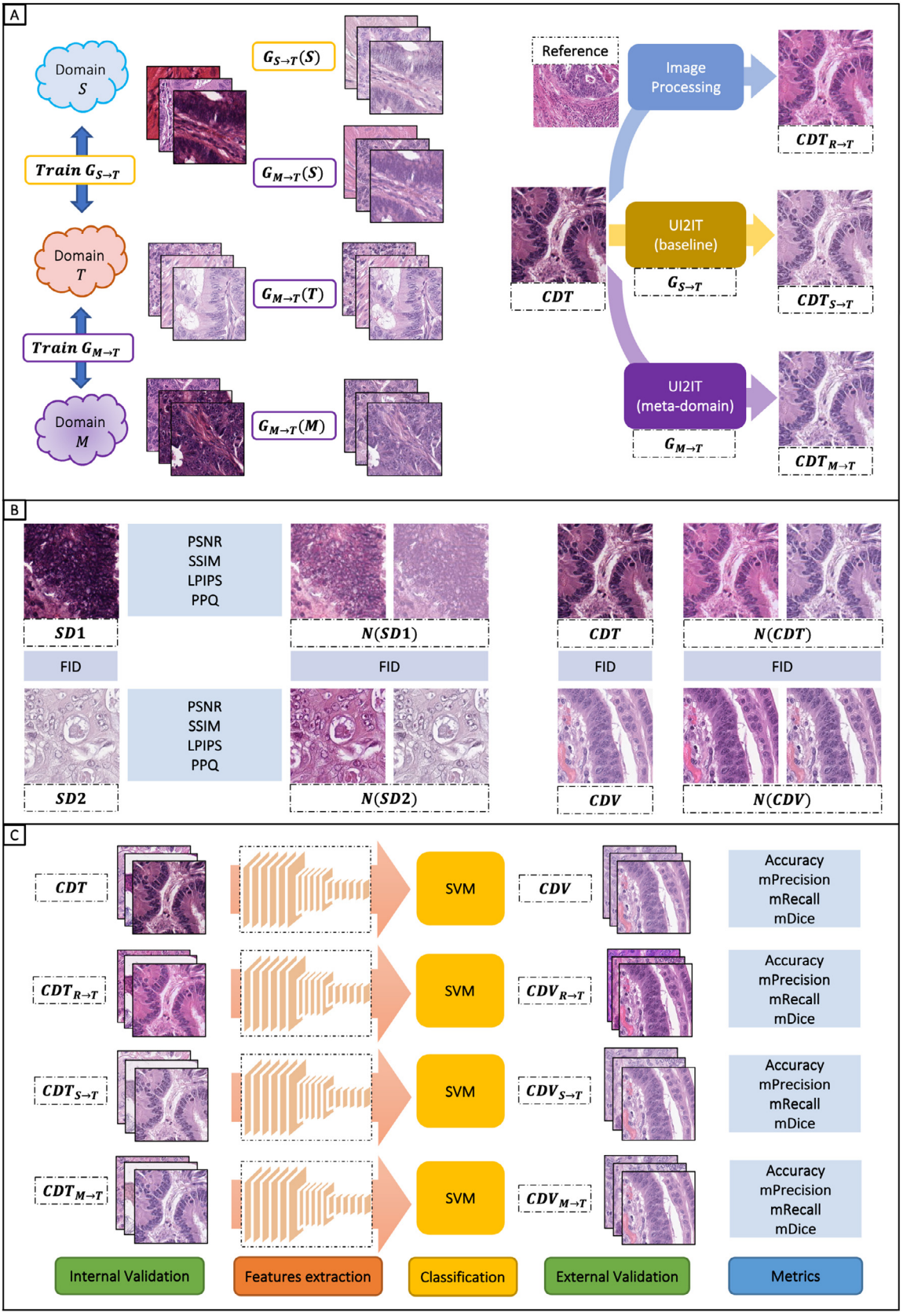

Fig. 3. Workflow employed for the study. (A) Training and use of the stain color normalization module. The GANs belonging to the UI2IT framework are trained by exploiting a meta-domain *M*. Source and Target domains, *S* and *T*, refer to the training dataset and the external validation dataset for classification, respectively. Traditional color normalization techniques exploit a reference tile *R*. (B) Assessment of the stain color normalization models. (C) Training and validation of the CRC tissue classifier. The classifiers are trained on features separately extracted for every normalization technique.

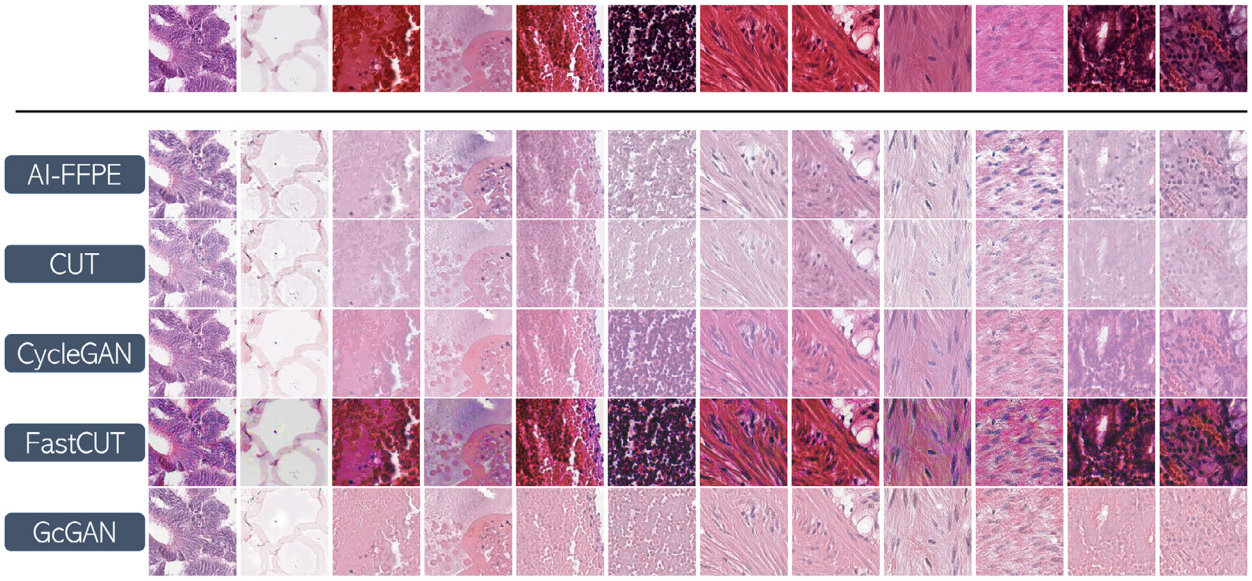

Fig. 4. Normalization examples with the various UI2IT GANs considered.

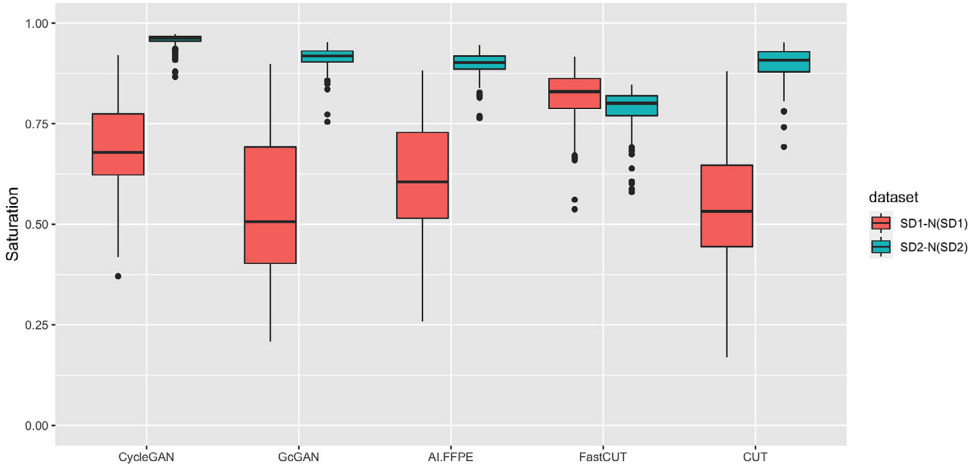

Fig. 5. Boxplots for SSIM between saturation channel of normalized images versus original one.

scribes the quantitative relationships between stain color normalization metrics and classification performances.

4.1. Stain color normalization

A visual example of the different techniques based on UI2IT for stain color normalization is portrayed in Fig. 4. It is possible to notice different color patterns for the various methods. In particular, FastCUT tends to generate tiles which have a higher saturation than the ones generated from other methods. This fact can be seen from Figs. 5 and 6. Boxplots in Fig. 5 shows that, for the $SD1$ dataset, $N(SD1)$ images obtained with FastCUT are the ones with saturation more similar to the original version. This is because FastCUT is the method which perform the lesser desaturation. Instead, for the $SD2$ dataset, $N(SD2)$ images obtained with FastCUT are the ones with the greatest dissimilarity to the original ones. This is because other methods do not saturate these images, instead FastCUT increases the saturation of input images.

Image similarity measures, namely PSNR, SSIM, and LPIPS, have been calculated between $SD1$ and $N(SD1)$ and $SD2$ and $N(SD2)$, in order to assess presence of artifacts or image degradation when performing stain color normalization. These quantitative results are reported in Table 1 and Fig. 7. In the comparison between $SD1$ and $N(SD1)$, we can note that, among the classical normalization methods, Macenko displayed the highest PSNR and SSIM, being 22.75 ± 5.17 and 0.95 ± 0.05, respectively, and the lowest LPIPS, being of 0.06 ± 0.05. For PSNR, the FastCUT method achieved better performances, with a value of 27.61 ± 1.41. For SSIM and LPIPS, the GAN-based methods obtained lower results than Macenko's method. In the comparison between $SD2$ and $N(SD2)$, we can observe that CycleGAN showed the best values for PSNR, SSIM, and LPIPS, being of 35.15 ± 1.73, 0.98 ± 0.01, and 0.05 ± 0.01, respectively.

Dissimilarity between distributions of features extracted from the images coming from the different domains are reported in Table 2 and Fig. 8. In the comparison on the stain normalization datasets, CycleGAN achieved the lowest FID, being 15.53, for the single normalization set-up. GcGAN, instead, displayed the best FID for the double normalization configuration, with a value of 15.65. On the classification datasets, with the single normalization, AI-FFPE obtained the best FID, being of 67.36. The results drastically improve with a double normalization, in which FastCUT achieved an FID of 49.60 on the classification datasets.

The PPQ metric has been used to assess the GAN-generated image quality from a pathologist perspective. It has been checked between source image tiles and corresponding generated normalized images, so that the pathologist can assess not only the reality of

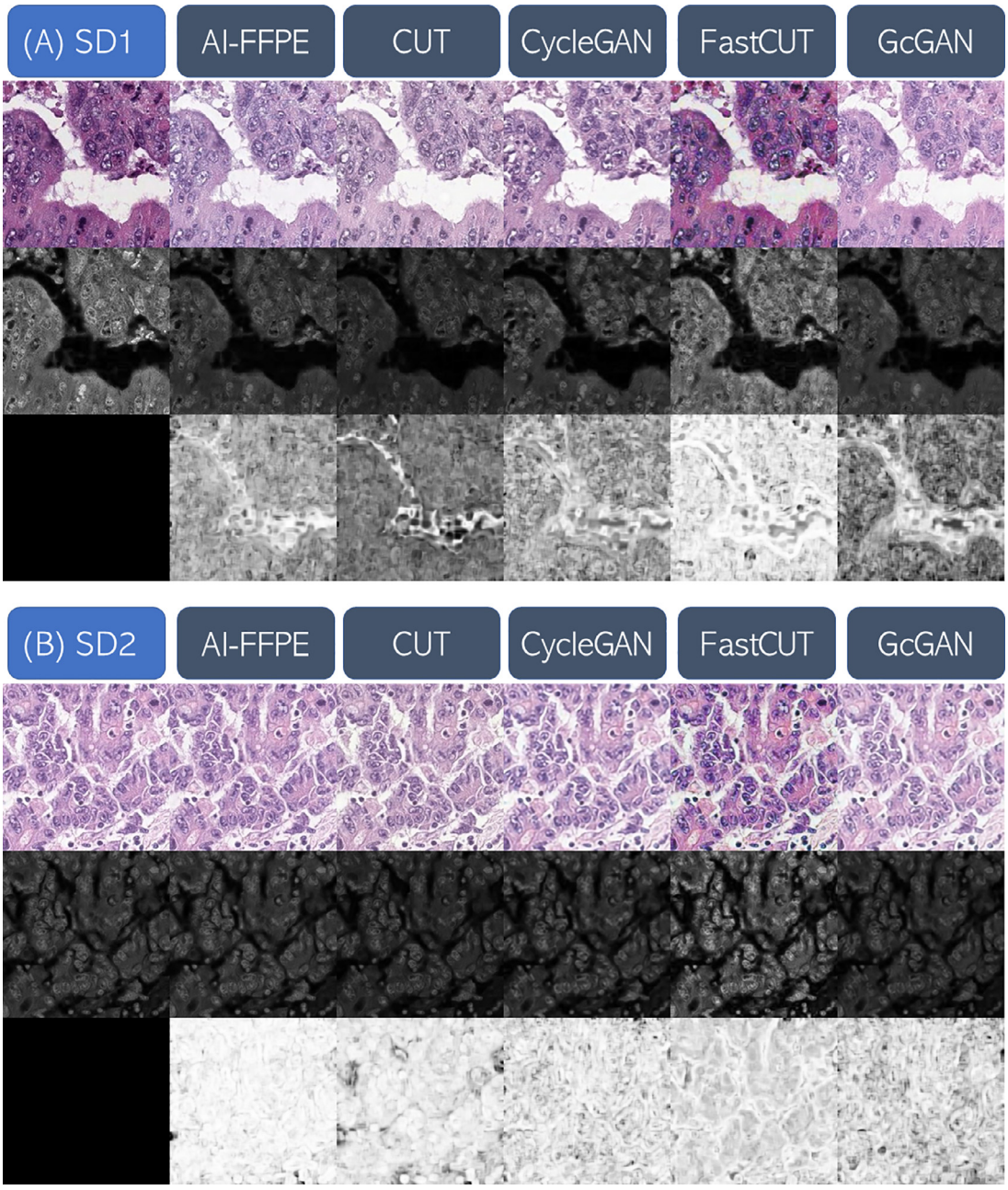

Fig. 6. Saturation differences between the considered UI2IT methods. (A) Tile from SD1 dataset. (B) Tile from SD2 dataset. (First Row) Original tile in the first column, then different GAN-based normalization methods. (Second Row) Saturation channel of the images in the first row. (Third Row) Difference between saturation of normalized image compared to the original one.

the generated image, but also its consistency with respect to the original one and the lack of artifacts. Quantitative results are portrayed in Table 3 and Fig. 9. CycleGAN demonstrated the best PPQ in the normalization of the SD1 and SD2 datasets, being the values of 1.39 ± 0.82 and 1.02 ± 0.21, respectively.

4.2. Multi-class classification

The accuracy of the multi-class classification models has been assessed both with internal cross-validation on the *CDT* dataset, and with external validation on the locally collected *CDV* dataset. These quantitative results are shown graphically in Fig. 10, and numerically in Tables 4 and 5. Other classification metrics are reported in Tables 6 and 7, for the traditional and UI2IT-based normalization methods, respectively. The DenseNet201 model consistently outperformed the other deep feature extractors in all the scenarios, with the only exception of the FastCUT model in the baseline UI2IT configuration. As shown in Table 4, amid the traditional normalization methods, the Reinhard is the one which allowed to obtain the highest validation accuracy. As reported in Table 5, among the GAN-based normalization methods, trained with the exploitation of the meta-domain, FastCUT achieved slightly better performances than the other methods. FastCUT also presented better performance than the Reinhard method. The lower results of FastCUT in the baseline configuration may be due to the fact that it does not correctly learn the saturation difference between the two domains, as reported in Section 3.5. On the other hand, the introduction of the meta-domain and the double normalization of the proposed approach compensate with the changes in the saturation values. In the base-

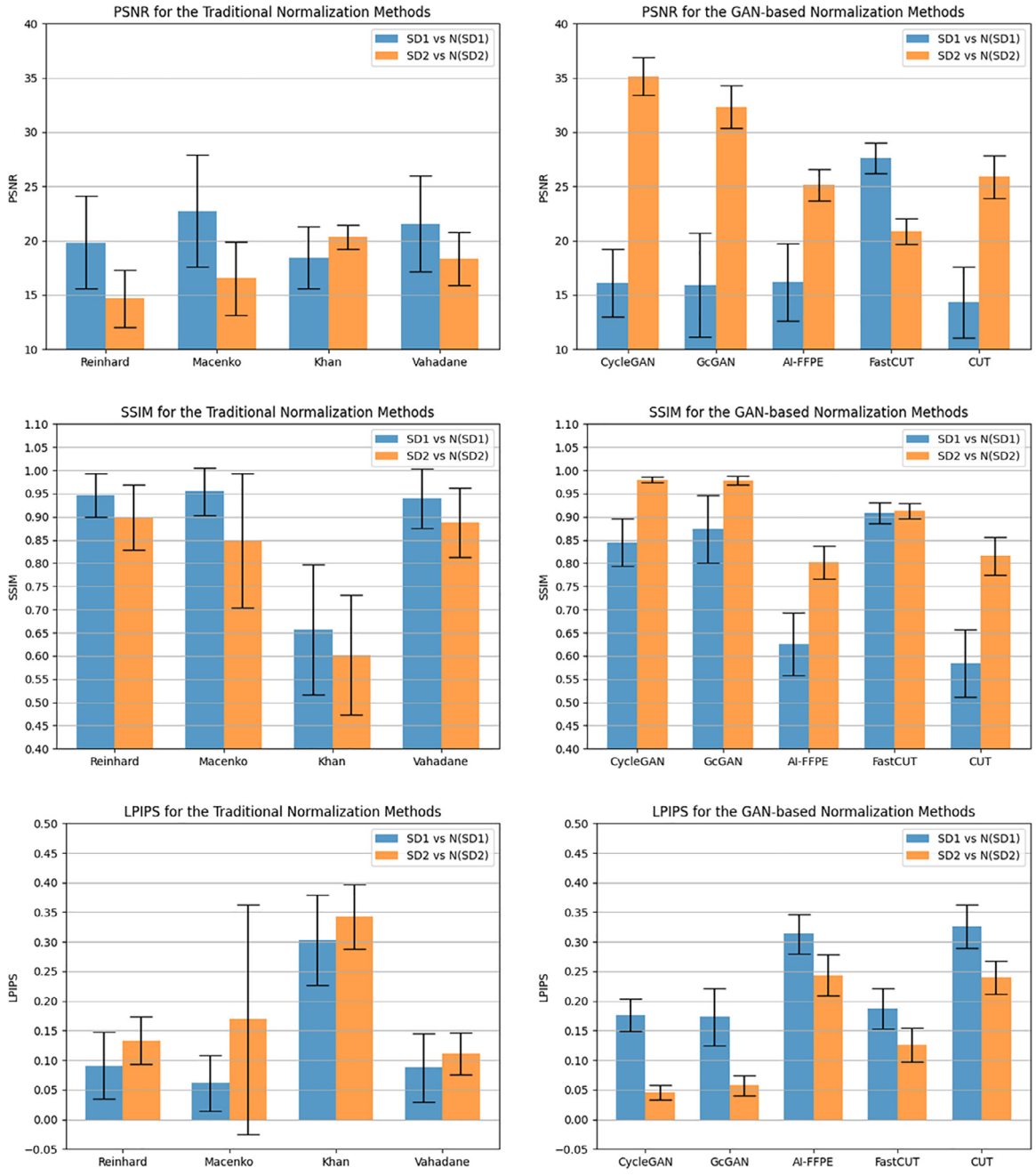

Fig. 7. Bar plots with error bars of PSNR, SSIM, LPIPS for N(SD1) vs SD1 and N(SD2) vs SD2.

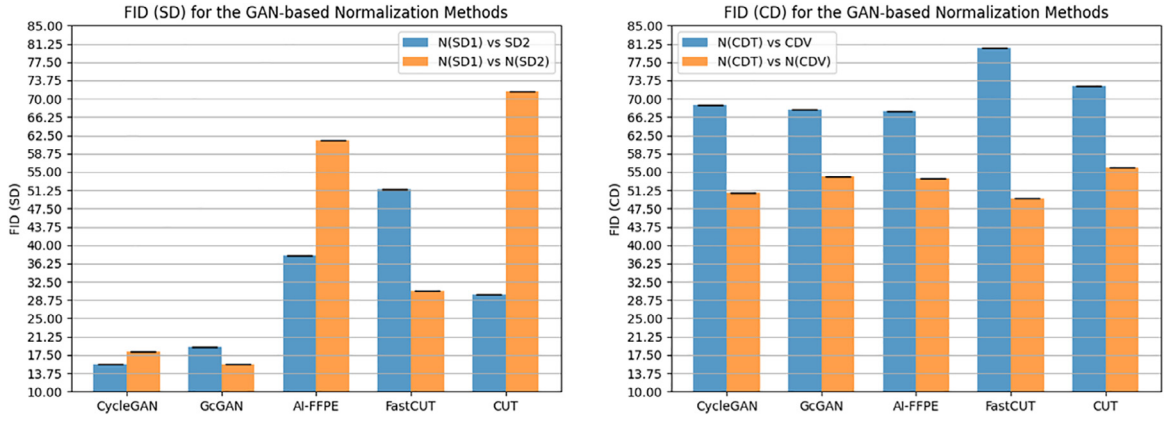

Fig. 8. Bar plots with error bars of FID for N(SD1) vs SD2, N(SD1) vs N(SD2), N(CDT) vs CDV, and N(CDT) vs N(CDV).

Table 1
Mean and standard deviation of PSNR, SSIM, LPIPS for N(SD1) vs SD1 and N(SD2) vs SD2.

Normalization	Data	PSNR ↑	SSIM ↑	LPIPS VGG ↓
Reinhard	N(SD1) vs SD1	19.8302 ± 4.2694	0.9462 ± 0.0462	0.0911 ± 0.0565
	N(SD2) vs SD2	14.6660 ± 2.6482	0.8992 ± 0.0702	0.1333 ± 0.0400
Macenko	N(SD1) vs SD1	22.7494 ± 5.1671	**0.9542 ± 0.0516**	**0.0615 ± 0.0474**
	N(SD2) vs SD2	16.5445 ± 3.3880	0.8491 ± 0.1449	0.1690 ± 0.1936
Khan	N(SD1) vs SD1	18.4453 ± 2.8344	0.6564 ± 0.1403	0.3028 ± 0.0761
	N(SD2) vs SD2	20.3537 ± 1.0981	0.6019 ± 0.1292	0.3424 ± 0.0542
Vahadane	N(SD1) vs SD1	21.5565 ± 4.4068	0.9397 ± 0.0637	0.0875 ± 0.0575
	N(SD2) vs SD2	18.3096 ± 2.4472	0.8875 ± 0.0747	0.1115 ± 0.0356
CycleGAN	N(SD1) vs SD1	16.0873 ± 3.1129	0.8449 ± 0.0508	0.1767 ± 0.0270
	N(SD2) vs SD2	**35.1468 ± 1.7279**	**0.9801 ± 0.0067**	**0.0455 ± 0.0127**
GcGAN	N(SD1) vs SD1	15.9142 ± 4.7748	0.8730 ± 0.0727	0.1731 ± 0.0487
	N(SD2) vs SD2	32.3358 ± 1.9626	0.9783 ± 0.0097	0.0577 ± 0.0172
AI-FFPE	N(SD1) vs SD1	16.2018 ± 3.5457	0.6255 ± 0.0672	0.3137 ± 0.0335
	N(SD2) vs SD2	25.1388 ± 1.4350	0.8019 ± 0.0350	0.2436 ± 0.0350
FastCUT	N(SD1) vs SD1	**27.6084 ± 1.4109**	0.9083 ± 0.0227	0.1874 ± 0.0336
	N(SD2) vs SD2	20.8744 ± 1.1850	0.9130 ± 0.0163	0.1256 ± 0.0286
CUT	N(SD1) vs SD1	14.3408 ± 3.2684	0.5837 ± 0.0730	0.3258 ± 0.0366
	N(SD2) vs SD2	25.8854 ± 1.9922	0.8157 ± 0.0406	0.2398 ± 0.0277

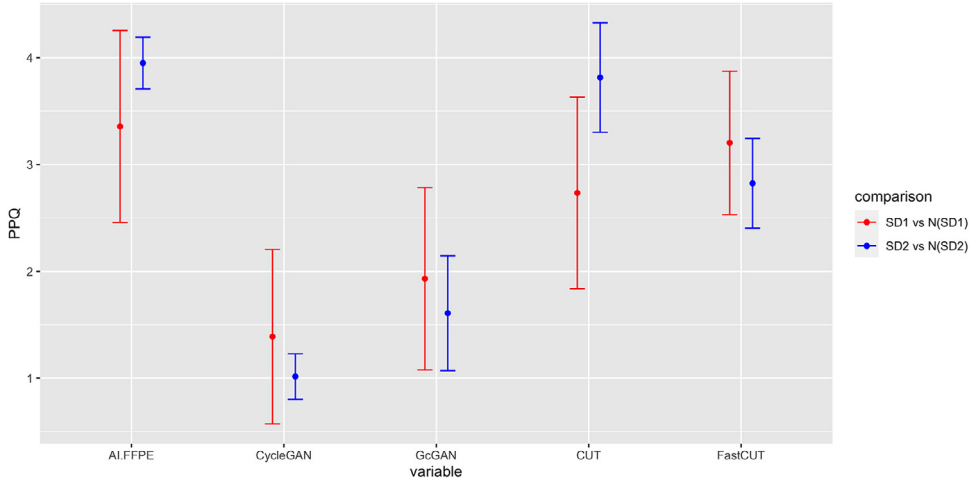

Fig. 9. Line Plots of PPQ for SD1 vs N(SD1) and SD2 vs N(SD2).

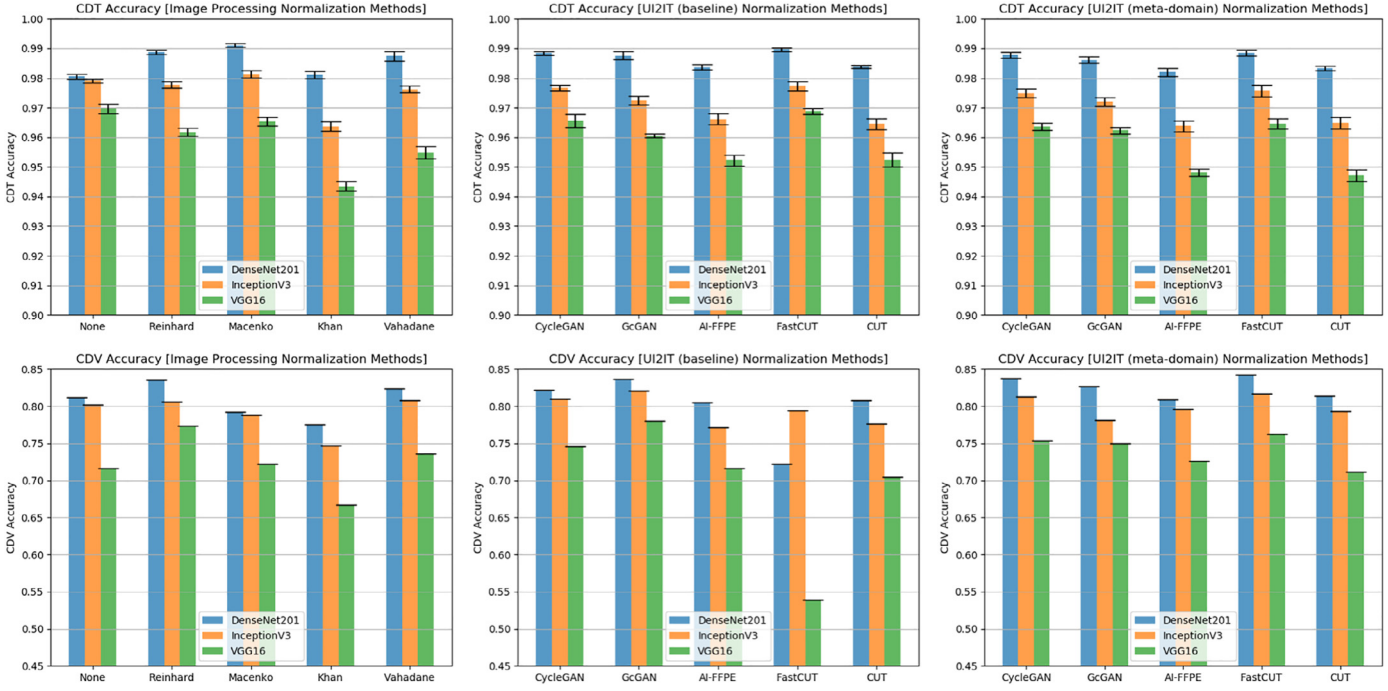

Fig. 10. Accuracy for the various normalization methods on the CDT (first row) and CDV (second row) datasets. First column: image processing normalization. Second column: UI2IT (baseline) normalization. Third column: UI2IT (meta-domain) normalization.

Table 2
FID comparison for the various normalization methods.

Normalization	Data	FID ↓
None	SD1 vs SD2	35.9791
	CD1 vs CD2	73.8733
Reinhard	N(SD1) vs N(SD2)	26.6528
	N(CD1) vs N(CD2)	58.1405
Macenko	N(SD1) vs N(SD2)	26.6980
	N(CD1) vs N(CD2)	53.7302
Khan	N(SD1) vs N(SD2)	22.9181
	N(CD1) vs N(CD2)	44.1937
Vahadane	N(SD1) vs N(SD2)	26.6518
	N(CD1) vs N(CD2)	56.6086
CycleGAN	N(SD1) vs SD2	**15.5333**
	N(SD1) vs N(SD2)	18.2566
	N(CD1) vs CD2	68.6784
	N(CD1) vs N(CD2)	50.7333
GcGAN	N(SD1) vs SD2	19.2417
	N(SD1) vs N(SD2)	**15.6546**
	N(CD1) vs CD2	67.8117
	N(CD1) vs N(CD2)	54.1416
AI-FFPE	N(SD1) vs SD2	37.8711
	N(SD1) vs N(SD2)	61.5486
	N(CD1) vs CD2	**67.3626**
	N(CD1) vs N(CD2)	53.6819
FastCUT	N(SD1) vs SD2	51.5162
	N(SD1) vs N(SD2)	30.5891
	N(CD1) vs CD2	80.4216
	N(CD1) vs N(CD2)	**49.5950**
CUT	N(SD1) vs SD2	29.9900
	N(SD1) vs N(SD2)	71.4490
	N(CD1) vs CD2	72.5915
	N(CD1) vs N(CD2)	55.9412

Table 3
Mean and standard deviation of PPQ for N(SD1) vs SD1 and N(SD2) vs SD2.

Normalization Method	Data	PPQ ↓
CycleGAN	N(SD1) vs SD1	**1.3883 ± 0.8165**
	N(SD2) vs SD2	**1.0151 ± 0.2127**
GcGAN	N(SD1) vs SD1	1.9309 ± 0.8531
	N(SD2) vs SD2	1.6080 ± 0.5385
AI-FFPE	N(SD1) vs SD1	3.3564 ± 0.8990
	N(SD2) vs SD2	3.9497 ± 0.2410
FastCUT	N(SD1) vs SD1	3.2021 ± 0.6715
	N(SD2) vs SD2	2.8241 ± 0.4195
CUT	N(SD1) vs SD1	2.7340 ± 0.8977
	N(SD2) vs SD2	3.8141 ± 0.5131

line setup, the GcGAN model showed the best performance, with results slightly higher than GcGAN trained with the meta-domain configuration, but lower than FastCUT in the meta-domain setup.

The quality of the features has been assessed by looking at embedding plots for the normalization methods. The embedding plot associated to the features extracted by DenseNet201, after FastCUT normalization, in the meta-domain configuration, can be seen from Fig. 11.

4.3. Relationships between normalization and classification

Several interesting relationships have been observed by our analysis. Indeed, as it can be observed from Fig. 12, there is a trend between performance of the classification methods and quality of the distribution of the generated images. This trend is visible for all the considered metrics, but it is statistically significant only for Top3-Accuracy (VGG16 achieves $r = 0.9446$, $p = 0.0155$), mDice (VGG16 achieves $r = 0.9075$, $p = 0.0333$), and mPrecision (DenseNet201 achieves $r = 0.9950$, $p = 0.0004$ and InceptionV3 achieves $r = 0.9495$, $p = 0.0135$). FID has been normalized by taking its reciprocal, so to make it correspond to a measure in which higher values correspond to better quality distributions.

The PPQ has also been found correlated to FID and to LPIPS, as can be seen from Fig. 13. These results are also statistically significant (PPQ SD1 vs FID: $r = 0.8968$, $p = 0.0392$; PPQ SD2 vs LPIPS SD2: $r = 0.9775$, $p = 0.0040$).

5. Discussion

In this study, we characterized several GAN-based UI2IT methods with the aim to perform stain color normalization for the classification of CRC histopathological tissue. Contrarily to what is usually done for GAN-based normalization, that is, considering only source and target domains, we added a meta-domain, which in our case consists in WSIs belonging to the TCGA. This meta-domain, containing tissues from a wide variety of laboratories, can allow to learn a mapping to our target domain that is more general than previous research works, thus avoiding the need to train a normalization model for every pair of possible data domains for histopathological data.

5.1. Stain color normalization

To assess the distribution of the GAN-generated images, FID has been assessed in different configurations. In the traditional UI2IT setting, one is expected to translate images from source to target domain. Then, the distribution of translated images is compared

Table 4
Accuracy on the CDT and CDV classification datasets using the classical normalization methods.

Normalization	Architecture	CDT	CDV		
		Acc	Top 1-Acc	Top 2-Acc	Top 3-Acc
None	InceptionV3	0.9791	0.8025	0.9209	0.9608
	DenseNet201	0.9805	0.8115	0.9253	**0.9721**
	VGG16	0.9697	0.7168	0.8688	0.9376
Reinhard	InceptionV3	0.9777	0.8063	0.9133	0.9586
	DenseNet201	0.9888	**0.8352**	0.9261	0.9652
	VGG16	0.9617	0.7737	0.8964	0.9473
Macenko	InceptionV3	0.9813	0.7884	0.8978	0.9486
	DenseNet201	**0.9910**	0.7924	0.8969	0.9557
	VGG16	0.9653	0.7219	0.8701	0.9244
Khan	InceptionV3	0.9637	0.7465	0.8871	0.9449
	DenseNet201	0.9811	0.7757	0.9070	0.9560
	VGG16	0.9435	0.6673	0.8319	0.9168
Vahadane	InceptionV3	0.9762	0.8075	0.9179	0.9603
	DenseNet201	0.9874	0.8240	**0.9265**	0.9646
	VGG16	0.9548	0.7362	0.8787	0.9390

Table 5
Accuracy on the CDT and CDV classification datasets using the UI2IT normalization methods. The methods are trained on both the map between the meta-domain and the target dataset ($G_{M \to T}$) and the source domain and the target dataset ($G_{S \to T}$).

Model	Architecture	$G_{M \to T}$				$G_{S \to T}$			
		CDT	CDV			CDT	CDV		
		Acc	Top 1-Acc	Top 2-Acc	Top 3-Acc	Acc	Top 1-Acc	Top 2-Acc	Top 3-Acc
CycleGAN	InceptionV3	0.9749	0.8126	0.9256	0.9678	0.9767	0.8097	0.9244	0.9668
	DenseNet201	0.9878	0.8375	**0.9413**	**0.9793**	0.9884	0.8213	0.9359	0.9780
	VGG16	0.9636	0.7538	0.8923	0.9461	0.9656	0.7463	0.8786	0.9420
GcGAN	InceptionV3	0.9720	0.7811	0.9161	0.9657	0.9725	0.8205	0.9333	0.9700
	DenseNet201	0.9861	0.8266	0.9363	0.9747	0.9876	**0.8370**	**0.9418**	0.9769
	VGG16	0.9623	0.7494	0.8841	0.9410	0.9605	0.7808	0.8972	0.9460
AI-FFPE	InceptionV3	0.9638	0.7957	0.9062	0.9536	0.9662	0.7713	0.9147	0.9692
	DenseNet201	0.9820	0.8090	0.9217	0.9693	0.9837	0.8053	0.9255	**0.9794**
	VGG16	0.9481	0.7263	0.8662	0.9355	0.9522	0.7166	0.8477	0.9103
FastCUT	InceptionV3	0.9757	0.8170	0.9305	0.9702	0.9774	0.7945	0.9247	0.9653
	DenseNet201	**0.9886**	**0.8420**	0.9410	0.9774	**0.9897**	0.7218	0.8733	0.9516
	VGG16	0.9646	0.7626	0.8956	0.9516	0.9689	0.5388	0.7097	0.8528
CUT	InceptionV3	0.9649	0.7935	0.9127	0.9615	0.9645	0.7767	0.8992	0.9608
	DenseNet201	0.9833	0.8134	0.9262	0.9728	0.9838	0.8075	0.9138	0.9735
	VGG16	0.9471	0.7111	0.8517	0.9332	0.9524	0.7049	0.8443	0.9116

Table 6
Classification mean precision, recall, and dice (mPrecision, mRecall, and mDice, respectively), using the traditional normalization methods.

Normalization	Architecture	CDV		
		mPrecision	mRecall	mDice
None	InceptionV3	0.8361	0.7668	0.7835
	DenseNet201	**0.8479**	0.7920	0.8039
	VGG16	0.8007	0.6743	0.6967
Reinhard	InceptionV3	0.8354	0.7780	0.7902
	DenseNet201	0.8643	**0.8038**	**0.8172**
	VGG16	0.8268	0.7278	0.7492
Macenko	InceptionV3	0.8171	0.7594	0.7722
	DenseNet201	0.8291	0.7740	0.7874
	VGG16	0.8291	0.6778	0.7203
Khan	InceptionV3	0.7893	0.7092	0.7263
	DenseNet201	0.8249	0.7397	0.7623
	VGG16	0.7289	0.6070	0.6218
Vahadane	InceptionV3	0.8369	0.7801	0.7948
	DenseNet201	0.8605	0.7955	0.8124
	VGG16	0.8219	0.6890	0.7226

to the distribution of images that originally belong to the target domain. In our setting, with the introduction of the meta-domain corresponding to TCGA images, we can instead observe that distribution of images that are normalized from both source and target classification domains are closer if compared to distributions obtained by performing only the normalization of the classification source domain.

Indeed, by performing a double normalization at inference time (that is, performing both $G_{MT}(S)$ and $G_{MT}(T)$, instead of only $G_{MT}(S)$), an improvement in the FID among the data distributions of the classification dataset can be observed. Fig. 8 shows that FID is consistently lower for all the considered GAN models when normalization is performed to both CDT and CDV, and not only CDT. This may look counterintuitive, since the learned mapping has already WSIs of the target domain coming from the same distribution as CDV. Instead, when considering the stain color normalization datasets, we note that performing double normalization is not always convenient. Indeed, SD1 and SD2 are the source and target domain, respectively, adopted for learning the GAN mapping. In this case, CycleGAN, AI-FFPE, and CUT show worse FID results when double normalization is applied instead of single normalization.

In this study, we introduced PPQ, a perceptive quality metric based on the assessment performed by the pathologist on the GAN-generated image tiles. As can be seen from Fig. 9, CycleGAN and GcGAN were the two best-performing methods according to PPQ. This result is perfectly consistent with FID values observed

Table 7
Classification mean precision, recall, and dice (mPrecision, mRecall, and mDice, respectively), UI2IT normalization methods. The methods are trained on both the map between the meta-domain and the target dataset ($G_{M \to T}$) and the source domain and the target dataset ($G_{S \to T}$).

Normalization	Architecture	$G_{M \to T}$			$G_{S \to T}$		
		CDV			CDV		
		mPrecision	mRecall	mDice	mPrecision	mRecall	mDice
CycleGAN	InceptionV3	0.8543	0.7754	0.7952	0.8112	0.7527	0.7714
	DenseNet201	0.8761	0.8151	0.8349	0.8460	0.7896	0.8110
	VGG16	0.8263	0.7153	0.7384	0.7552	0.6841	0.6908
GcGAN	InceptionV3	0.8399	0.7439	0.7684	0.8679	0.7802	0.8041
	DenseNet201	0.8645	0.8030	0.8229	**0.8802**	**0.8104**	**0.8347**
	VGG16	0.8187	0.7130	0.7324	0.8168	0.7467	0.7627
AI-FFPE	InceptionV3	0.8470	0.7542	0.7754	0.8112	0.7527	0.7714
	DenseNet201	0.8664	0.7788	0.8003	0.8460	0.7896	0.8110
	VGG16	0.8105	0.6927	0.7158	0.7552	0.6841	0.6908
FastCUT	InceptionV3	0.8667	0.7784	0.8017	0.8570	0.7497	0.7747
	DenseNet201	**0.8800**	**0.8170**	**0.8372**	0.7955	0.6999	0.7099
	VGG16	0.8069	0.7315	0.7522	0.6868	0.5184	0.5047
CUT	InceptionV3	0.8397	0.7614	0.7823	0.8064	0.7516	0.7667
	DenseNet201	0.8616	0.7858	0.8067	0.8361	0.7931	0.8097
	VGG16	0.7829	0.6865	0.7085	0.7533	0.6712	0.6808

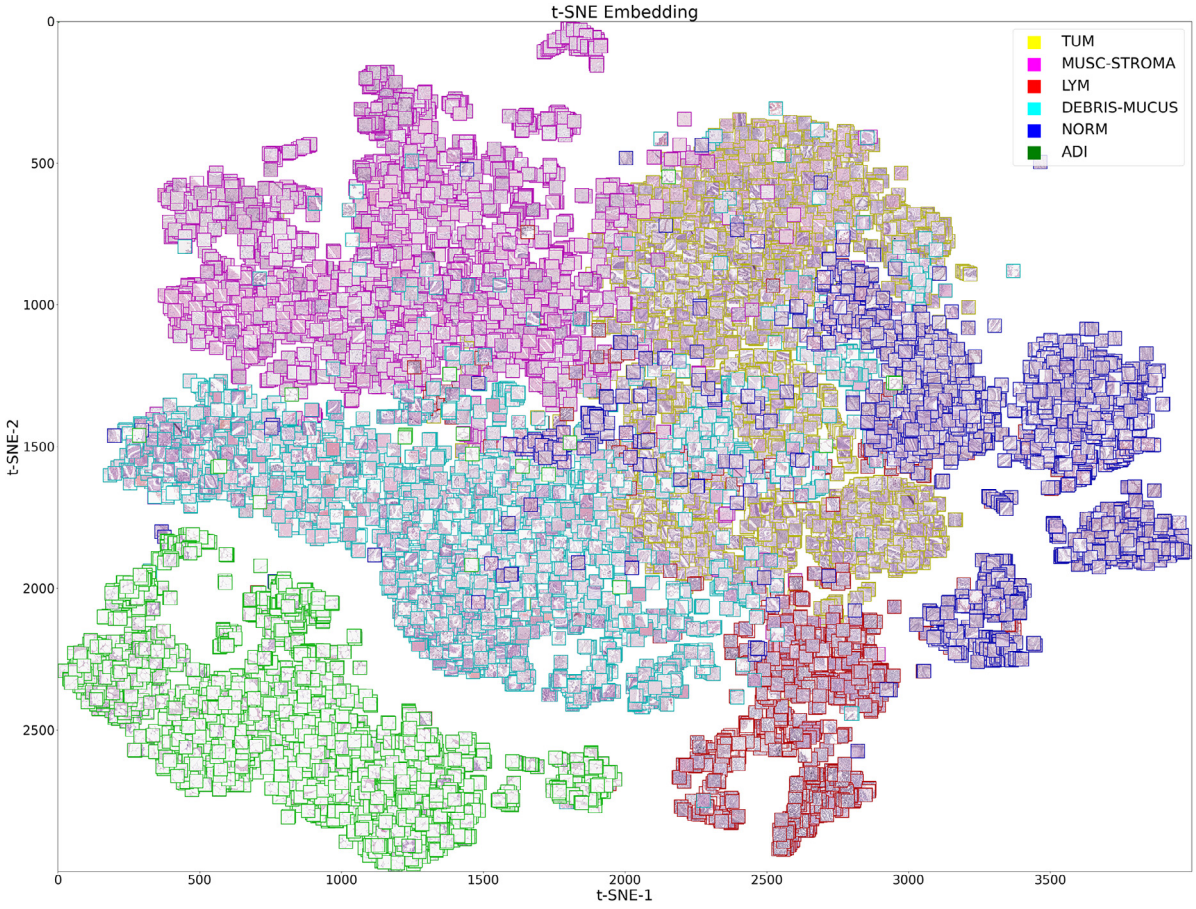

Fig. 11. Embedding plot obtained applying t-SNE to features extracted from DenseNet201 on the CDV dataset, with the FastCUT normalization methodology.

on $N(SD1)$ vs $SD2$ and $N(SD1)$ vs $N(SD2)$, and with LPIPS values observed on $SD1$ vs $N(SD1)$ and $SD2$ vs $N(SD2)$. Therefore, this study can confirm that CNN-features-based metrics to assess the perceptive quality of generated images can be exploited also in histopathological contexts. As can be seen from Fig. 13, FID and PPQ, and LPIPS and PPQ, are also correlated when considering the GAN-based normalization results.

The effectiveness of CycleGAN and GcGAN images is also observed by the fact that they possess the highest values for PSNR and SSIM for what concerns $SD2$ vs $N(SD2)$, as observed in Fig. 7. This means that those methods are the most suitable for performing transformation of the target domain images with $G_{MT}(T)$, which is useful as previously mentioned (considering the increase in FID for double normalization).

5.2. Multi-class classification

With respect to the multi-class classification task, the DenseNet201 model consistently outperforms InceptionV3 and VGG16, in all considered normalization paradigms, as portrayed in Fig. 10. The only anomaly is for the FastCUT model in the baseline UI2IT configuration, for which Inception V3 displays the best performance. The normalization methodology which allows to achieve the best performance on the locally collected external validation set is FastCUT in the meta-domain configuration. Though FastCUT does not achieve satisfactory results in reality of generated image tiles, since it tends to over-saturate the input images (quantitative assessment, see Fig. 5; visual examples, see Fig. 6), when applied in double normalization configuration, it can still offer a useful normalization for the classification task. This is a behavior which would not be possible in the baseline configuration considered for stain color normalization with UI2IT, in which normalization is accomplished only from source to target domain. It is worth noting the comparison between FID of $N(SD1)$ vs $SD2$ with FID of $N(SD1)$ vs $N(SD2)$, and FID of $N(CDT)$ vs CDV with FID of $N(CDT)$ vs $N(CDV)$. Indeed, after double normalization is performed, FastCUT achieves the best FID on the classification datasets. This result means that perceptive quality of transformed images is not necessarily correlated to better classification accuracies of downstream classifiers. Though, we consider perceptive quality of generated images a useful asset for the introduction of GAN-based normalization models in the clinical routine.

The quality of the features extracted by DenseNet201 after having performed normalization with FastCUT, in the meta-domain configuration, can be seen from Fig. 11. The different tissue classes appear clearly clustered in the 2D embedding plot obtained with t-SNE.

Classification results are correlated to the quality of image distributions, as can be seen from correlations between normalized FID and classification performance reported in Fig. 12. This confirms that FID or similar quality metrics should be checked when performing GAN-based normalization (i.e., conditional generation), since they can already offer an important insight on how the models adopted for the downstream tasks will perform.

5.3. Limitations

Although the results of the proposed framework are promising, leading to better generalization capabilities and higher performance, without the need to train a style transfer model be-

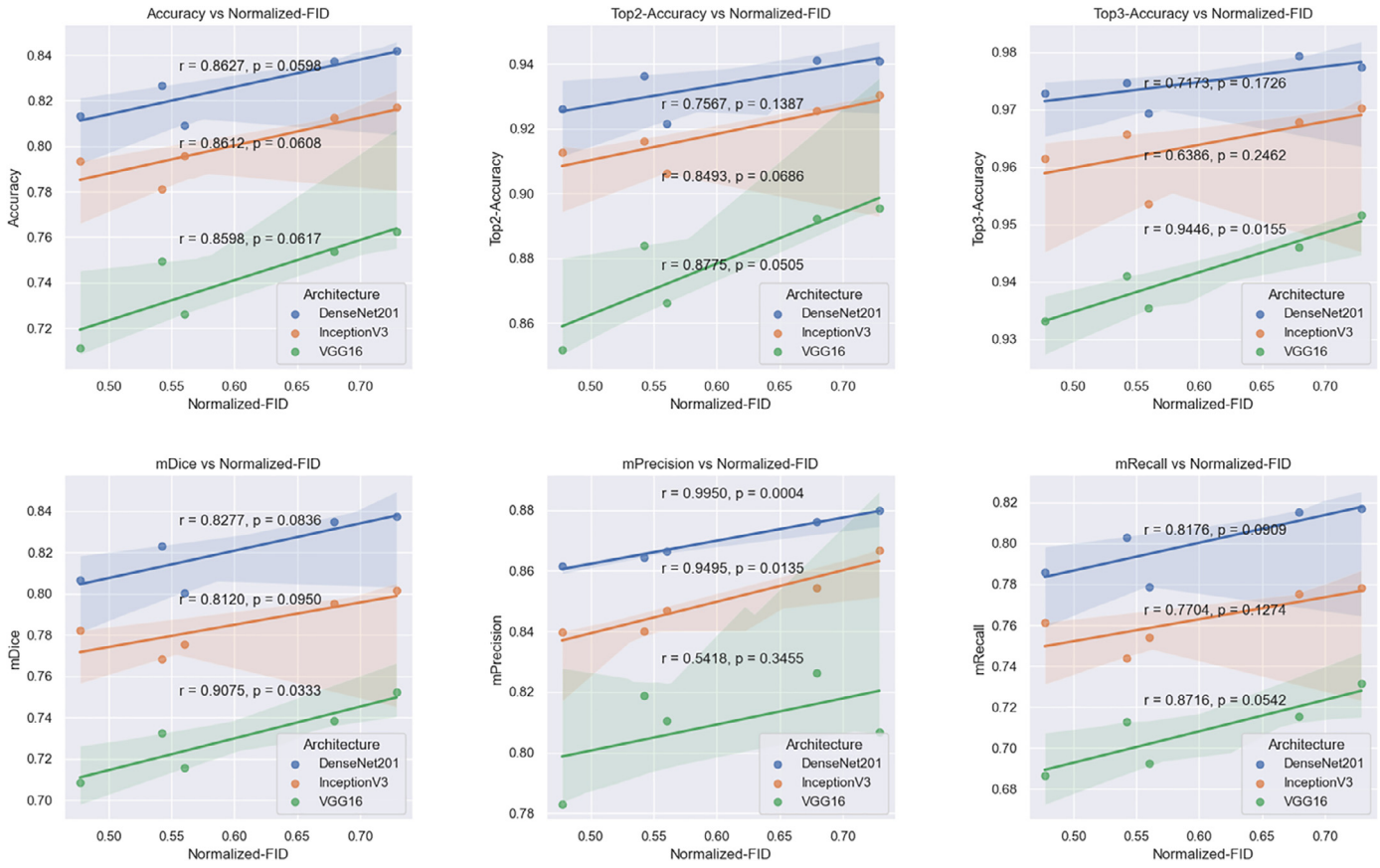

Fig. 12. Correlations between classification metrics and FID.

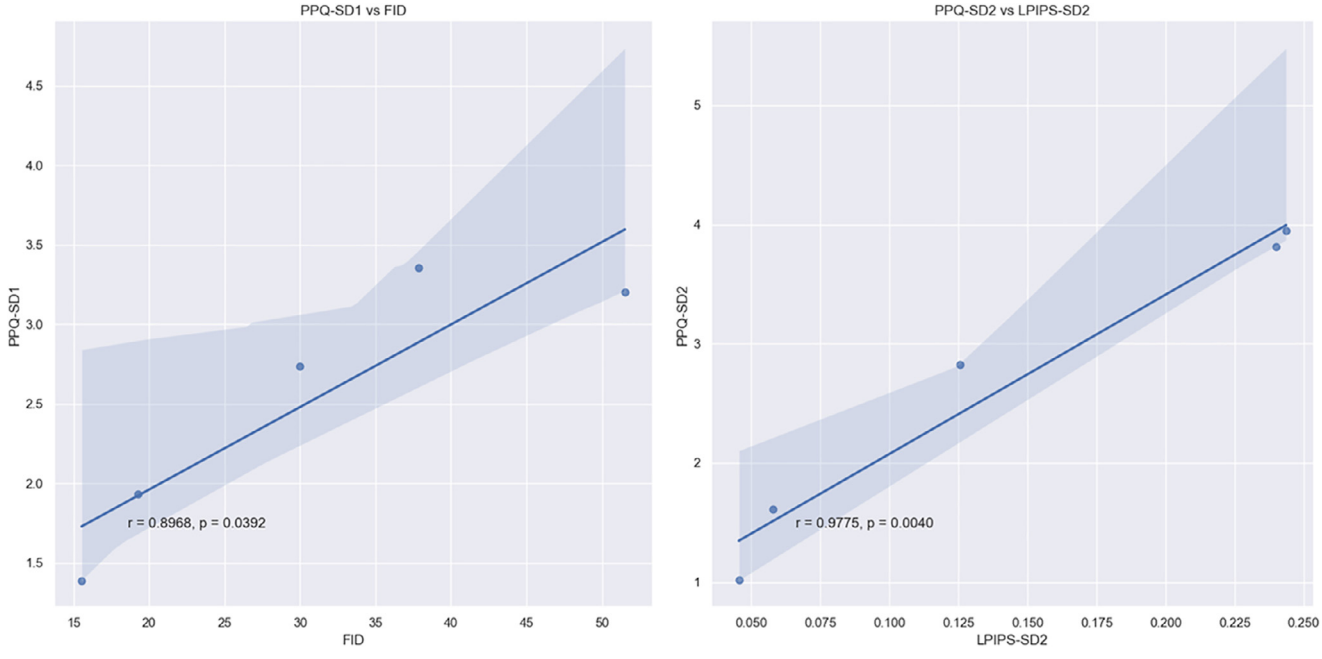

Fig. 13. Correlation between FID and PPQ, and between LPIPS and PPQ.

tween each pair of laboratories, there are still points that can be improved in future studies.

One limitation is that other downstream tasks can be considered to further show the validity of the proposed pipeline in other contexts, such as nuclei segmentation [61]. Indeed, a more comprehensive analysis of downstream tasks would allow realizing a general stain transfer paradigm that can greatly aid quantitative pipelines for Digital Pathology environments.

Another limitation is that the proposed metric for assessing reality of generated images from a pathologist perspective, PPQ; has been assessed by a single expert. In the future, a collaborative quantitative measure, which comprehends the evaluation of more

pathologists, can be introduced to provide a more objective quantification of GAN-generated image tiles in Digital Pathology setups.

6. Conclusion

In this work, three CNN architectures and nine normalization techniques have been considered for the sake of realizing a pipeline which is robust to stain color variation in a CRC histological classification setup.

CRC multi-class tissue classification is an important task in digital pathology. In particular, to date, the study of tumors is moving toward the integration of genomic data, such as transcriptomics and its spatial localization. In the present paper, we focused on CRC, but the described approach could be considered as a proof of concept for other malignancies. Indeed, the segmentation task could be a preliminary step in digital pathology studies, for instance, to study the relationship between the tumor, its microenvironment, and genomic features.

Since preparation of histological slides is a complex process, composed of various stages, and differences can be introduced in any of them, color normalization is a fundamental step to effectively perform quantitative tasks.

Traditional normalization methods are easy-to-use, since they only need a template tile, but generate images which are subject to color artifact, making them unsuitable for subsequent pathologists' analyses. On the other side, the realm of GAN architectures to achieve UI2IT can offer a powerful framework to carry out stain color normalization, allowing to achieve impressive results both from the quality and reality of generated tiles, and for the performance of the models involved in the subsequent tasks.

Nonetheless, care has to be reserved for assessing GAN-generated image tiles, and several quality measures should be included to have a complete overview of which model may work better for the task under consideration.

Observations of pathologists may help in assessing quality of generated images. Indeed, we included them in our study with the PPQ metric, which can be used to have a medical expert view that can eventually confirm the quality of the distributions measured with quantitative measures as FID.

The introduction of a meta-domain during the learning phase of the stain transfer model, as proposed by this study, can help reduce the training time of normalization models for a specific laboratory, and also provide better generalization capabilities of downstream classifiers trained after normalization.

Author contributions

NA, TMM, SDS drafted the paper; NA, VB conceived the experiments; TMM, MC, NA, BP performed the experiments; FAZ provided the internal datasets; ST, AA, AB, VB, BP, MC revised the manuscript; SDS, VB supervised the project; EM annotated the tiles and provided the PPQ scores; FAZ, EM revised generated images.

Funding

This research has been funded by the projects:

- "D3 4 Health – Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care", project code: PNC0000001, Concession Decree No. 931 of 6 June 2022 adopted by the Italian Ministry of University and Research, CUP: B53C22006170001, funded under the National Plan for National Recovery and Resilience Plan (NRRP) Complementary Investments – Law Decree May 6, 2021, n. 59, converted and modified as to Law n. 101/2021 Research initiatives for technologies and innovative trajectories in the health and care sectors – Italian Ministry of University and Research funded by the European Union – NextGenerationEU;

- "Tecnopolo per la Medicina di Precisione", CUP: B84I18000540002;

- Italian Ministry of Health "Ricerca Corrente 2022".

Ethical statement

The Institutional Ethics Committee of the IRCCS Istituto Tumori "Giovanni Paolo II" approved the study (Prot n. 780/CE).

The authors affiliated to the IRCCS Istituto Tumori "Giovanni Paolo II", Bari, are responsible for the views expressed in this article, which do not necessarily represent the Institute.

Declaration of Competing Interest

The authors declare that they have no competing interests.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2020, CA Cancer, J. Clin. 70 (1) (2020) 7–30, doi:10.3322/caac.21590.

[2] J. Gao, Z. Shen, Z. Deng, L. Mei, Impact of tumor–stroma ratio on the prognosis of colorectal cancer: a systematic review, Front. Oncol. 11 (2021) 738080, doi:10.3389/fonc.2021.73808011.

[3] T.A.A. Tosta, P.R. de Faria, L.A. Neves, M.Z. do Nascimento, Computational normalization of H&E-stained histological images: progress, challenges and future potential, Artif. Intell. Med. 95 (2019) 118–132.

[4] M. Salvi, U.R. Acharya, F. Molinari, K.M. Meiburger, The impact of pre-and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis, Comput. Biol. Med. 128 (2021) 104129.

[5] C.M. Chen, Y.S. Huang, P.W. Fang, C.W. Liang, R.F. Chang, A computer-aided diagnosis system for differentiation and delineation of malignant regions on whole-slide prostate histopathology image using spatial statistics and multidimensional densenet, Med. Phys. 47 (3) (2020) 1021–1033.

[6] F. Ciompi, O. Geessink, B.E. Bejnordi, G.S. De Souza, A. Baidoshvili, G. Litjens, J Van Der Laak, The importance of stain normalization in colorectal tissue classification with convolutional networks, in: Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017, pp. 160–163. IEEE.

[7] F.G. Zanjani, S. Zinger, B.E. Bejnordi, J.A. van der Laak, P.H. de With, Stain normalization of histopathology images using generative adversarial networks, in: Proceedings of the 2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018), 2018, pp. 573–577. IEEE.

[8] J.T. Pontalba, T. Gwynne-Timothy, E. David, K. Jakate, D. Androutsos, A. Khademi, Assessing the impact of color normalization in convolutional neural network-based nuclei segmentation frameworks, Front. Bioeng. Biotechnol. 7 (2019) 300.

[9] M. Runz, D. Rusche, S. Schmidt, M.R. Weihrauch, J. Hesser, C.A. Weis, Normalization of HE-stained histological images using cycle consistent generative adversarial networks, Diagn. Pathol. 16 (1) (2021) 1–10.

[10] Z. Swiderska-Chadaj, T. de Bel, L. Blanchet, A. Baidoshvili, D. Vossen, J. van der Laak, G. Litjens, Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer, Sci. Rep. 10 (1) (2020) 1–14.

[11] H. Cho, S. Lim, G. Choi, H. Min, Neural stain-style transfer learning using GAN for histopathological images, arXiv preprint 11 (2017).

[12] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[13] T. de Bel, J.M. Bokhorst, J. van der Laak, G. Litjens, Residual cyclegan for robust domain transformation of histopathological tissue slides, Med. Image Anal. 70 (2021) 102004.

[14] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graph. Appl. 21 (5) (2001) 34–41.

[15] M. Macenko, M. Niethammer, J.S. Marron, D. Borland, J.T. Woosley, X. Guan, N.E. Thomas, A method for normalizing histology slides for quantitative analysis, in: Proceedings of the 2009 IEEE international symposium on biomedical imaging: from nano to macro, 2009, pp. 1107–1110. IEEE.

[16] A.M. Khan, N. Rajpoot, D. Treanor, D. Magee, A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution, IEEE Trans. Biomed. Eng. 61 (6) (2014) 1729–1738.

[17] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, N. Navab, Structure-preserving color normalization and sparse stain separation for histological images, IEEE Trans. Med. Imaging 35 (8) (2016) 1962–1971.

[18] B.E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, J.A Van Der Laak, Stain specific standardization of whole-slide histopathological images, IEEE Trans. Med. Imaging 35 (2) (2015) 404–415.

[19] Nicola Altini, Tommaso Maria Marvulli, Francesco Alfredo Zito, Mariapia Caputo, Stefania Tommasi, Amalia Azzariti, Antonio Brunetti, Berardino Prencipe, Eliseo Mattioli, Simona De Summa, Vitoantonio Bevilacqua, Colorectal cancer histology image tiles for tissue multi-class classification [data set], Zenodo (2022), doi:10.5281/zenodo.7109754.
[20] N. Linder, J. Konsti, R. Turkki, E. Rahtu, M. Lundin, S. Nordling, J. Lundin, Identification of tumor epithelium and stroma in tissue microarrays using texture analysis, Diagn Pathol 7 (1) (2012) 1–11.
[21] R.M. Haralick, K. Shanmugam, I.H. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. (6) (1973) 610–621.
[22] V. Bevilacqua, N. Pietroleonardo, V. Triggiani, A. Brunetti, A.M. Di Palma, M. Rossini, L. Gesualdo, An innovative neural network framework to classify blood vessels and tubules based on Haralick features evaluated in histological images of kidney biopsy, Neurocomputing 228 (2017) 143–153.
[23] J.N. Kather, C.A. Weis, F. Bianconi, S.M. Melchers, L.R. Schad, T. Gaiser, F.G. Zöllner, Multi-class texture analysis in colorectal cancer histology, Sci. Rep. 6 (1) (2016) 1–11.
[24] N. Altini, T.M. Marvulli, M. Caputo, E. Mattioli, B. Prencipe, G.D. Cascarano, F.A. Zito, Multi-class tissue classification in colorectal cancer with handcrafted and deep features, in: Proceedings of the International Conference on Intelligent Computing, 2021, pp. 512–525. Springer, Cham.
[25] J.N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.A. Weis, N. Halama, Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study, PLoS Med. 16 (1) (2019) e1002730.
[26] A. BenTaieb, G. Hamarneh, Adversarial stain transfer for histopathology image analysis, IEEE Trans. Med. Imaging 37 (3) (2017) 792–802.
[27] J. Ke, Y. Shen, X. Liang, D. Shen, Contrastive learning based stain normalization across multiple tumor in histopathology, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021, pp. 571–580. Springer, Cham.
[28] T. Kausar, A. Kausar, M.A. Ashraf, M.F. Siddique, M. Wang, M. Sajid, I. Riaz, SA-GAN: stain acclimation generative adversarial network for histopathology image analysis, Appl. Sci. 12 (1) (2021) 288.
[29] Jakob Nikolas Kather, Image tiles of TCGA-CRC-DX histological whole slide images, non-normalized, tumor only (v0.1) [data set], Zenodo (2020), doi:10.5281/zenodo.3784345.
[30] Jakob Nikolas Kather, Niels Halama, Alexander Marx, 100,000 histological images of human colorectal cancer and healthy tissue (v0.1) [Data set], Zenodo (2018), doi:10.5281/zenodo.1214456.
[31] M.T. Shaban, C. Baur, N. Navab, S. Albarqouni, Staingan: stain style transfer for digital histological images, in: Proceedings of the 2019 IEEE 16th international symposium on biomedical imaging (Isbi 2019), IEEE, 2019, pp. 953–956.
[32] D. Bug, S. Schneider, A. Grote, E. Oswald, F. Feuerhake, J. Schüler, D. Merhof, in: Context-based Normalization of Histological Stains Using Deep Convolutional features. In Deep Learning in Medical Image Analysis and Multimodal Learning For Clinical Decision Support, Springer, Cham, 2017, pp. 135–142.
[33] M. Salvi, F. Molinari, U.R. Acharya, L. Molinaro, K.M. Meiburger, Impact of stain normalization and patch selection on the performance of convolutional neural networks in histological breast and prostate cancer classification, Comput. Methods Programs Biomed. Update 1 (2021) 100004.
[34] V. Sandfort, K. Yan, P.J. Pickhardt, R.M. Summers, Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks, Sci. Rep. 9 (1) (2019) 1–9.
[35] Y. Chen, Y. Zhao, W. Jia, L. Cao, X. Liu, Adversarial-learning-based image-to-image transformation: a survey, Neurocomputing 411 (2020) 468–486.
[36] A. Alotaibi, Deep generative adversarial networks for image-to-image translation: a review, Symmetry 12 (10) (2020) 1705.
[37] Y. Pang, J. Lin, T. Qin, Z. Chen, Image-to-image translation: methods and applications, IEEE Trans. Multimed. (2021).
[38] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: a review, Med. Image Anal. 58 (2019) 101552.
[39] H. Huang, P.S. Yu, C. Wang, An introduction to image synthesis with generative adversarial nets, arXiv preprint 39 (2018).
[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[41] K. Kurach, M. Lučić, X. Zhai, M. Michalski, S. Gelly, A large-scale study on regularization and normalization in GANs, in: Proceedings of the International Conference on Machine Learning, 2019, pp. 3581–3590. PMLR.
[42] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[43] T. Park, A.A. Efros, R. Zhang, J.Y. Zhu, Contrastive learning for unpaired image–to-image translation, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 319–345. Springer, Cham.
[44] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2849–2857.
[45] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: Proceedings of the International Conference on Machine Learning, 2017, pp. 1857–1865. PMLR.
[46] M. Amodio, S. Krishnaswamy, Travelgan: image-to-image translation by transformation vector learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8983–8992.
[47] S. Benaim, L. Wolf, One-sided unsupervised domain mapping, Adv. Neural Inf. Process. Syst. 30 (2017).
[48] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, D. Tao, Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2427–2436.
[49] K.B. Ozyoruk, S. Can, B. Darbaz, et al., A deep-learning model for transforming the style of tissue images from cryosectioned to formalin-fixed and paraffin-embedded, Nat. Biomed. Eng. 6 (2022) 1407–1419, doi:10.1038/s41551-022-00952-9.
[50] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[51] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[52] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Adv. Neural Inf. Process. Syst. 29 (2016).
[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).
[54] D.C. Dowson, B. Landau, The Fréchet distance between multivariate normal distributions, J. Multivar. Anal. 12 (3) (1982) 450–455.
[55] L.N. Wasserstein, Markov processes on countable product space describing large systems of automata, Probl. Pered. Inform. 5 (1969) 64–73.
[56] C.X. Ren, A. Ziemann, J. Theiler, A.M. Durieux, Deep snow: synthesizing remote sensing imagery with generative adversarial nets, in: Proceedings of the IAlgorithms, Technologies, and Applications for Multispectral and Hyperspectral Imagery XXVI, 11392, SPIE, 2020, pp. 196–205.
[57] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization (2014) arXiv preprint. doi:10.48550/arXiv.1412.6980.
[58] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[59] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[60] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014) arXiv preprint. doi:10.48550/arXiv.1409.1556.
[61] N. Altini, A. Brunetti, E. Puro, M.G. Taccogna, C. Saponaro, F.A. Zito, V. Bevilacqua, NDG-CAM: nuclei detection in histopathology images with semantic segmentation networks and grad-CAM, Bioengineering 9 (9) (2022) 475.
[62] V.K. Morris, E.B. Kennedy, N.N. Baxter, A.B. Benson 3rd, A. Cercek, M. Cho, K.K. Ciombor, C. Cremolini, A. Davis, D.A. Deming, M.G. Fakih, S. Gholami, T.S. Hong, I. Jaiyesimi, K. Klute, C. Lieu, H. Sanoff, J.H. Strickler, S. White, J.A. Willis, C. Eng, Treatment of metastatic colorectal cancer: ASCO guideline, J. Clin. Oncol. 41 (3) (2023) 678–700 Jan 20Epub 2022 Oct 17. PMID: 36252154, doi:10.1200/JCO.22.01690.